SRRC-RR-65-94

AFCRL-65-826

EVALUATION OF SPEECH PROCESSING DEVICES

I. INTELLIGIBILITY, QUALITY, SPEAKER RECOGNIZABILITY

by

William D. Voiers
Marion F. Cohen
Juozas Mickunas

Contract No. AF19(628)4195
Project 4610 Task 461002
Final Report: 31 July 1965
Period Covered: 1 June 1964 — 31 July 1965

Prepared for

AIR FORCE CAMBRIDGE RESEARCH LABORATORIES
OFFICE OF AEROSPACE RESEARCH
UNITED STATES AIR FORCE
BEDFORD, MASSACHUSETTS

S R R C

AFCRL-65-826 SRRC-RR-65-94

EVALUATION OF SPEECH PROCESSING DEVICES

I. INTELLIGIBILITY, QUALITY, SPEAKER RECOGNIZABILITY

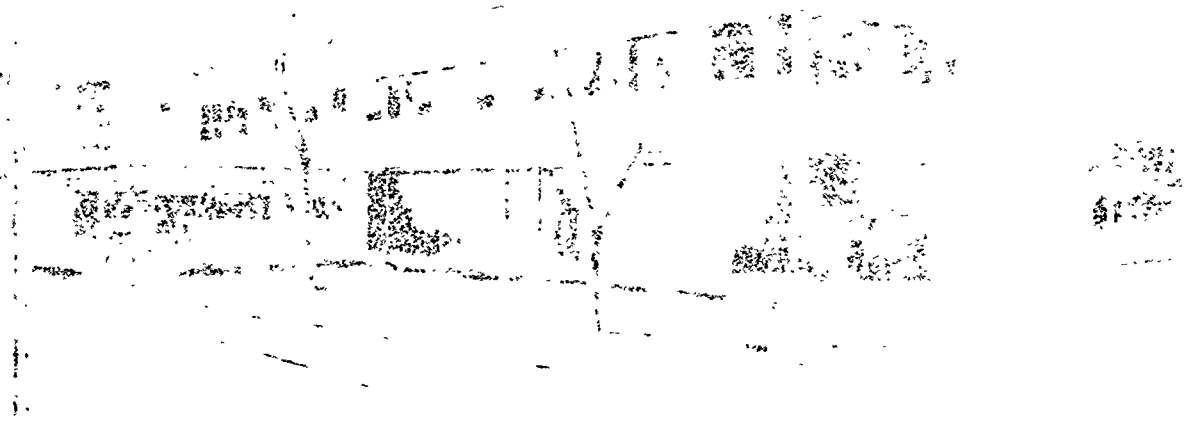
by

William D. Voiers
Marion F. Cohen
Juozas Mickunas

Contract No. AF19(628)4195
Project 4610 Task 461002
Final Report: 31 July 1965
Period Covered: 1 June 1964 — 31 July 1965

Prepared for

AIR FORCE CAMBRIDGE RESEARCH LABORATORIES
OFFICE OF AEROSPACE RESEARCH
UNITED STATES AIR FORCE
BEDFORD, MASSACHUSETTS



SPERRY RAND RESEARCH CENTER
SUDBURY, MASS.

ABSTRACT

This study is concerned with the development of improved methods of evaluating experimentally processed speech and, in turn, speech-processind devices and systems. Three bases of evaluation are dealt with in the study. These are: Intelligibility, Speaker Recognizability and Aesthetic Acceptability or Quality.

A two-choice diagnostic rhyme test for the transmission of consonant information has been developed. It yields a total intelligibility score plus diagnostic scores relating to the fidelity with which seven binary attributes of consonant phonemes are transmitted to the ear of the listener. These attributes are voicing, nasality, duration and frication (as opposed to plosion) i.e., front (as opposed to middle) middle (as opposed to back) and back (as opposed to front).

For treating the problem of speaker recognizability, procedures have been developed by means of which listeners' ratings of voices on various perceived acoustic traits can be analyzed to predict speaker recognizability under any given transmission condition.

The problem of evaluating the aesthetic acceptability or quality of transmitted speech is treated by means of the standard unit-variance method. Here, primary emphasis is placed upon the contributions of the channel to the quality of the received speech. However, the method is adaptable for purposes of studying qualitative variation attributable to the source (i.e., the speaker). In this method, speech as processed by four representative vocoder systems provides standards with which experimentally processed speech is compared by listeners. Listener response data are analyzed to yield a value representing the position of the experimental system on a standard unit-variance scale of aesthetic acceptability.

Results of evaluations of representative vocoders are presented for each of the three evaluation methods.

TABLE OF CONTENTS

TABLE OF CONTENTS (cont.)

LIST OF FIGURES

LIST OF FIGURES (cont.)

LIST OF TABLES

INTRODUCTION

The research described in this report is concerned with three distinct, though inter-related, problem areas in the field of speech perception: Speech Intelligibility, Speech Quality, and Speaker Recognizability. In each of these areas the research is necessarily concerned in some degree not only with fundamental principles of speech perception, but also with the reduction of these principles to practical techniques for evaluating the performance of modern communication systems and devices.

More than forty distinct experiments were conducted during the time period covered by this program. Approximately half of the total research effort which these involved was directly supported by government funds under Contract No. AF19(628)-4195. The remaining effort was supported by the Sperry Rand Corporation. In no instance, however, was the relevance of a particular experiment or series of experiments contingent upon source of support. Moreover, several experiments conducted under contract with non-governmental agencies also yielded a number of methodological insights, even though the results were of no intrinsic value in relation to the major aims of the program.

From the foregoing, it is perhaps apparent that a comprehensive account of procedures and results for all relevant research within the body of the present report would tend inevitably to jeopardize continuity in the treatment of the major issues.

The organization of this report was designed, as far as possible, to achieve completeness with minimum sacrifice of continuity and clarity.

A chapter is devoted to each of the major problem areas of intelligibility, quality, and speaker recognizability. Only those experimental details which bear directly upon a subject under discussion are included in Chapters 1, 2 and 3. However, a set of experimental summaries contained in Appendix III serves to document the essential details of procedure and results of each experiment. These include not only experiments which fulfilled the primary purposes for which they were designed, but also experiments which, because of experimental error, equipment malfunction, or other reasons, failed to serve their primary purposes. Virtually all of the experiments conducted yielded data of value for one purpose or another. Those which failed in the objective of evaluating a particular experimental treatment did, in any case, provide useful data bearing upon such issues as reliability, effects of learning, speaker effects, listener idiosyncrasies, etc.

Some remarks are in order by way of providing background for the chapters devoted to the three major aspects of the program.

In the area of intelligibility, several approaches were investigated. One of these was the "free conversation" approach. Here, two or more listeners participate in a simulated communications task with or without constraints upon the permissible vocabulary. Various aspects of the resulting conversation and performance may be scored to yield an evaluation of the speech transmission or processing systems involved. Two experiments were concerned with this general approach. The first was conducted under the supervision of our consultant, Dr. T.B. Roby, of Tufts University, and is being published as a Sperry Rand Research Report. This study was concerned primarily with the methodological problem of devising a communications "task" suitable for purposes of free conversation

testing. The second investigation explored the feasibility of diagnostic scoring of speech reception in a two-person, simulated communications situation. In both cases the results obtained were sufficiently favorable to warrant additional research. However, further work in this area was not undertaken in the present program in order that priority could be given to the development of the Diagnostic Rhyme Test. Chapter 1 is concerned entirely with the development and uses of this test.

In the area of quality evaluation, several approaches were investigated. The first of these involved a variation of the iso-preference method developed by Munson and Karlin (1962). Briefly, this approach involves the scaling of experimentally processed speech in terms of "equivalent speech-to-noise ratio" for unprocessed speech.

An attractive feature of this approach is its use of a "physical yardstick" for quality measurement. However, preliminary studies revealed the listener's task to be extremely difficult due to the qualitative differences between the reference condition (clear speech in noise) and the typical experimental condition (vocoderized speech). Practical considerations of economy in the use of listener time also contributed to the decision of approaching the problem in a somewhat different way. The ultimate consequence of this decision was the unit variance method of quality evaluation, which is the primary subject of Chapter 2. Among other innovations incorporated into the unit-variance method is the use of vocoderized speech itself as a standard for the comparative evaluation.

Chapter 3 presents the background and development of the voice rating approach to the evaluation of speaker recognizability. While this approach was the primary concern of this aspect of the program, some effort

was devoted to the development of an alternative approach which could serve, among other things, the function of validating the voice rating approach. This alternative took the form of a multiple-choice speaker recognition test, and tapes for use with this test were prepared for a sample of sixteen voices. However, administration of the test proved to be so cumbersome and time-consuming that a decision was made to postpone indefinitely any attempt at experimental evaluation. This may be undertaken at some future date.

None of the results presented in these chapters should be construed to represent the ultimate performance capabilities of any of the vocoders involved. Many of these vocoders were used at very early stages in the course of their development. Some were deliberately used in a state of major or minor malfunction in order to achieve a desired degree or form of speech degradation. However, a special supplement is devoted to the presentation of results of evaluation of the AFCRL Polymodal Vocoder during the final weeks of the program. While still not representative of the ultimate performance of the Polymodal Vocoder in the modes evaluated, these results at least serve to document the characteristics of this vocoder at various stages of its development.

More than thirty diagnostic intelligibility evaluations were carried out for various modes of the AFCRL Polymodal Vocoder. A total of sixteen evaluations of output-speech quality were conducted while fourteen evaluations of potential speaker recognizability were performed.

The results of these various evaluations are, in general, of no permanent intrinsic interest, though they served as the basis for various engineering decisions made in the course of the development of the Polymodal Vocoder. These results were communicated verbally and/or by monthly letter report during the course of the program. They are not,

therefore, treated in the present report except as they bear upon some aspect of the methodology of communication-system evaluation. A supplementary report contains the results of some of the more important of these evaluations.

Six papers based on research conducted under this program were read at the Spring, 1965, meetings of the Acoustical Society of America, while three more were presented at the Fall meetings of the society.

CHAPTER 1

SYSTEM EVALUATION FROM THE STANDPOINT OF SPEECH INTELLIGIBILITY

DEVELOPMENT OF A DIAGNOSTIC RHYME TEST

A point of departure for the design of the Diagnostic Rhyme Test was provided by the taxonomy or system of classification developed for consonant sounds by Miller and Nicely (22). While not exhaustive, the system permits unique characterization of 16 consonant sounds in terms of five "features" or attributes of the articulatory process. The manner in which sounds are produced is, of course, only one of several possible bases of classification. Where it is desired in particular that the set of classificatory parameters employed correspond to experimentally independent parameters of system performance, some question may arise as to the appropriateness of a taxonomy developed on this basis. However, an examination of some of the potential alternatives reveals a number of equivalences. For example, a very nearly equivalent taxonomy can be derived from the set of distinctive features formulated by Jakobson and Halle (17). Here, ostensibly, classification is on the basis of perceived characteristics of speech sounds. Moreover, at least some of the parameters of this taxonomy have fairly well-defined physical acoustical correlates.

Thus, any issue as to the optimal basis of classification is likely to be of a more academic than practical consequence at our present level of understanding of the psychophysics of speech.

The articulatory features or attributes distinguished by Miller and Nicely are:

1. Voicing
2. Nasality
3. Affrication[1]
4. Duration
5. Place of Production

The attributes of <u>voicing</u>, <u>nasality</u> and <u>affrication</u> are intrinsically binary and are so treated by Miller and Nicely. However, more levels of varia-

[1] The term "affrication" is used, somewhat incorrectly, by Miller and Nicely in reference to the fricative-plosive opposition. We have retained this nomenclature in the interest of continuity. However, the more correct nomenclature will be employed in subsequent reports.

tion are potentially distinguishable in the case of duration and place of production. Miller and Nicely recognize two levels of duration, while they treat place of production as a ternary attribute.

For the present purposes, the ternary characterization of place of production was discarded in favor of a three-dimensional binary characterization: i.e., Front (as opposed to Middle); Middle (as opposed to Back); and Back (as opposed to Front). Although this somewhat arbitrary modification of the Miller and Nicely scheme was motivated primarily by practical considerations, some amount of theoretical justification can be found for it. The front/middle (P_{12}) opposition turns out to be the equivalent of the "grave/acute" feature opposition of Jakobson and Halle, while middle/back (P_{23}) and back/front (P_{31}) parallel (though in opposite directions) the compact/diffuse opposition of the "distinctive feature" system of classification. The latter two parallels raise the possibility of some amount of redundancy in the present system of consonant classification. However, the theoretical basis of this possibility did not seem sufficiently strong to warrant the combination of middle/back and back/front into a single taxonomic parameter without benefit of experimental confirmation of their equivalence.

A seven-dimensional binary taxonomy thus served as the basis for the development of the Diagnostic Rhyme Test (DRT). The primary function of this test is evaluating voice communications systems or devices in terms of phonemic information transmitted via seven binary attributes of consonant sounds.

The performance of listeners in discriminating the source states of the various attributes serves then as the basis of speech evaluation - and in turn system evaluation - as accomplished with the Diagnostic Rhyme Test.

Speech Materials

The Diagnostic Rhyme Test is composed of 112 rhyming word-pairs. The words comprising each pair differ only with respect to the initial consonant sound, more specifically with respect to a single binary attribute of consonant sounds. An example is the pair, "zeal-seal" in which the attribute voicing is present and absent respectively. Accordingly, the 112 rhyming word-pairs can be classified into seven categories on the basis of the distinguishing consonant attribute. Each of these groups can, in turn, be subclassified on the basis of vowel "region". Within each group, then, are four word pairs associated with each of four regions which are selected, so far as is practicable, to "bracket" the vowel triangle. These four regions are identified by the vowels [ɔ], [u], [i] and [e]. With one exception all of the words involved can be classified as CVC, CV, or CVCC. This exception, the word "thread", was necessary simply because the restrictions employed in the selection of speech materials so severely limited the population of acceptable English words. It was necessary, for the same reason, to use certain word pairs twice in order to obtain the required number of pairs for each attribute-vowel category. Table 1.1 shows the word pairs which are actually contained in each category.

From the structure of the speech materials used, it is perhaps apparent that listener-response data obtained with the DRT can be evaluated or scored in several ways. Of primary interest are scores relating to the discriminability of the seven critical attributes. By appropriate organization of listener response data it is possible to derive a gross measure of system performance with respect to each attribute. Further, for each of the seven attributes, a score may be obtained for that attribute when it is present in the stimulus word, and another score for words in which the attribute in question is not present in the stimulus word. The mean of these two scores then is a measure of system performance for a particular attribute. Thus, three scores are obtained for each of the seven

Table 1.1 Diagnostic Rhyme Test Materials

	/e/	/i/	/u/	/ɔ/
Voicing *(lax-tense)	bet - pet vend - fend den - ten dent - tent	zeal - seal dee - tea deem - team beak - peak	dunes - tunes dues - twos dune - tune do - to	gall - call vault - fault galled - called vaults - faults
Nasality (nasal-oral)	neck - deck knell - dell melt - belt mend - bend	kneel - deal neap - deep meat - beat me - bee	moot - boot nude - dude noose - deuce moor - boor	morn - born morn - born mauled - bald mall - ball
Affrication (Continuant-interrupted)	felt - pelt fend - pend thence - dense vent - bent	feel - peel feet - peat field - peeled feast - pieced	foo - pooh fools - pools fool - pool fooled - pooled	fawn - pawn fawn - pawn fall - pall fall - pall
Duration (strident-mellow)	zen - then zen - then said - thread said - thread	seem - theme seem - theme seems - themes seems - themes	sue - thew sue - thew sues - thews sues - thews	saw - thaw sawed - thawed saws - thaws sought - thought
Place of Articulation Front vs. Middle (grave - acute)	met - net bed - dead pence - tense pest - test	mead - need peach - teach fief - thief bean - dean	poor - tour moo - new boom - doom moon - noon	for - thor maws - gnaws maw - gnaw maud - gnawed
Middle vs. Back (diffuse-compact)	debt - get tend - kenned self - shelf sell - shell	seen - sheen teeth - keith teal - keel see - she	toot - coot suit - shoot tool - cool tomb - coom	sort - short torque - cork taught - caught daunt - gaunt
Back vs. Front (compact - diffuse)	guess - bess ken - pen keg - peg guest - best	key - pea keys - peas keep - peep keen - peen	ghoul - buhl goon - boon cooch - pooch goo - boo	caw - paw cawed - pawed cause - pause gawk - balk

* Equivalent "distinctive feature" after Cheny (1957).

attributes: 1) a gross score of responses to all words, 2) score of responses to words with attributes present, and 3) score of responses to words with attributes absent.

In addition to scoring the DRT with respect to the seven articulatory attributes, the test can also be scored to give a measure of consonant discriminability under various conditions of vowel context. Consonant discriminability scores can be obtained for each of the four vowels used in the DRT. In addition, each of the 21 "attribute scores" described above can be computed for each vowel context. It is thus possible to obtain a measure of consonant discriminability as a function of distinguishing attribute, vowel context and of their various combinations.

Although the DRT was not designed specifically to give a gross measure of system performance, an over-all intelligibility score can be derived. A more detailed discussion of this score and of evidence bearing on its validity is given below.

Correction for Effect of Chance

In obtaining any of the above-described intelligibility scores, a simple correction is made for the effect of chance by using the formula $\frac{(R-W)100}{T}$, where R is the number of correct responses, W is the number of incorrect responses, and T is the total number of stimulus words.

Forms of the Diagnostic Rhyme Test

In a complete DRT list of 224 words, each of the 112 word pairs is represented twice. The first word and every seventh word thereafter is selected from a word pair in which the initial consonants are alike, except that one is a voiced sound and the other an unvoiced sound (gall-call). The second word and every seventh word thereafter is selected from a word pair for which the critical attribute is nasality. The third and every seventh word thereafter tests for the discriminability of affrication and so on. Thus,

the test is designed so that an intelligibility score can easily be obtained for each of the seven articulatory dimensions.

The most frequently used form of the DRT is constructed so that the selection of the stimulus word from each pair is random except for the restriction that both source states of each attribute occur with equal frequency in each vowel context. Thus, every seventh word involves the same consonant attribute, but the stimulus word selected from each pair is randomly determined. Also, the order is random with respect to the four vowel sounds involved in each instance. There are four such lists of word pairs, and for each "random" list of word pairs there are four lists of word selections. Thus, there are 16 lists of 112 words. In each list of 112 words there are an equal number of words representing each attribute when it is present and when it is absent. Each vowel is also equally represented for each articulatory dimension. It is thus possible to create a balanced list of 224 words by selecting any two lists of 112 words.

There is, in addition to the random form of the DRT, a "single-attribute" form. The "single-attribute" form involves the 16 word pairs representing a particular dimension, or a total of 32 possible stimulus words. Vowel order is randomized, as is stimulus word selection. There are four such lists for each attribute.

For initial evaluation of the DRT, a special form of the test (the experimental form) was constructed. The order of articulatory dimensions from one to seven remains constant here, but the vowel order is not randomized. All of the word pairs representing a particular vowel are grouped together. Also, the stimulus words in the first half of the list (112 words) contain the word of each pair in which the attribute is present. The second half contains the word of each pair in which the attribute is absent. There are two such lists.

Each list was constructed for purposes of data analysis in the initial stages of DRT evaluation before computer-scoring techniques were implemented.

Speaker Selection

For purposes of initial evaluation of the DRT, a professionally trained speaker (RD) was selected to record the lists. Several other speakers representing extremes of certain perceptually significant voice characteristics, or perceived acoustic traits, have also been recorded and used in various studies. The speakers were judged to be as follows:

1. neutral (RC)
2. high-pitched
3. low-pitched
4. rough
5. smooth
6. clear
7. unclear

Preparation of Stimulus Materials

Recording procedures are described in Appendix IV. Initially the rate of stimulus presentation was one word every 2.5 seconds. Subsequently, however, a study was performed to determine the optimum rate of stimulus presentation (i.e., the fastest rate which produces a high reliability and a good subject performance). The rates investigated ranged from 0.7 second per word to 2.8 seconds per word. Based on the results of this study, a rate of one word each 1.33 seconds was adopted as a standard for all subsequent recordings. (Appendix II, Summaries I-3 and I-4). In all cases, the stimulus words were uttered without a carrier phrase.

Administration

In routine administration of the DRT, eight normal-hearing males between the ages of 17 and 30 are used as subjects. The response sheet consists of a list of word pairs, the order of which corresponds to the stimulus

word order of the particular list being used. The subjects are instructed to put a line through the word they hear. All materials are played on a high-quality recorder and presented binaurally over high-quality matched earphones at a vowel level of approximately 85 dB. Each new test condition is preceded by 50 practice words.

Reliability and Sensitivity

The DRT, recorded as described above for initial evaluation, was tested under various speech-to-noise conditions in order to demonstrate the sensitivity to the most common form of speech degradation. Subjects were 40 adult males who were employees of the Sperry Rand Research Center and had no previous exposure to the test. They were divided into five groups, each group consisting of eight listeners.

The stimulus material was presented binaurally over high-quality matched earphones. Each group of subjects listened to the entire test twice (448 words), half listening first to the recording of words in which the initial attribute was absent while the other half listened first to words in which the initial attribute was present. The level of the speech signal remained constant for all conditions. The noise and speech were band-passed at 60 Hz and 7500 Hz. The S/N ratios of -12 dB, -6 dB, 0 dB, +6 dB, and +12 dB were presented one to each group.

The results of this experiment, as shown in Fig. 1.1, indicate that over the range of speech-to-noise ratios from -12 dB to +12 dB the total DRT score has a gain function with a slope of approximately 3.5%/dB. As can be seen on the graph, this slope very closely approximates the results reported for the Fairbanks Rhyme Test (11). While the sensitivity of total scores tends to decrease at speech-to-noise ratios above +6 dB, various subtests retain a high degree of sensitivity up to and beyond +12 dB (Fig. 1.2).

Figure 1.2 also demonstrates the independent behavior of the various subtests under degraded conditions of speech. As can be

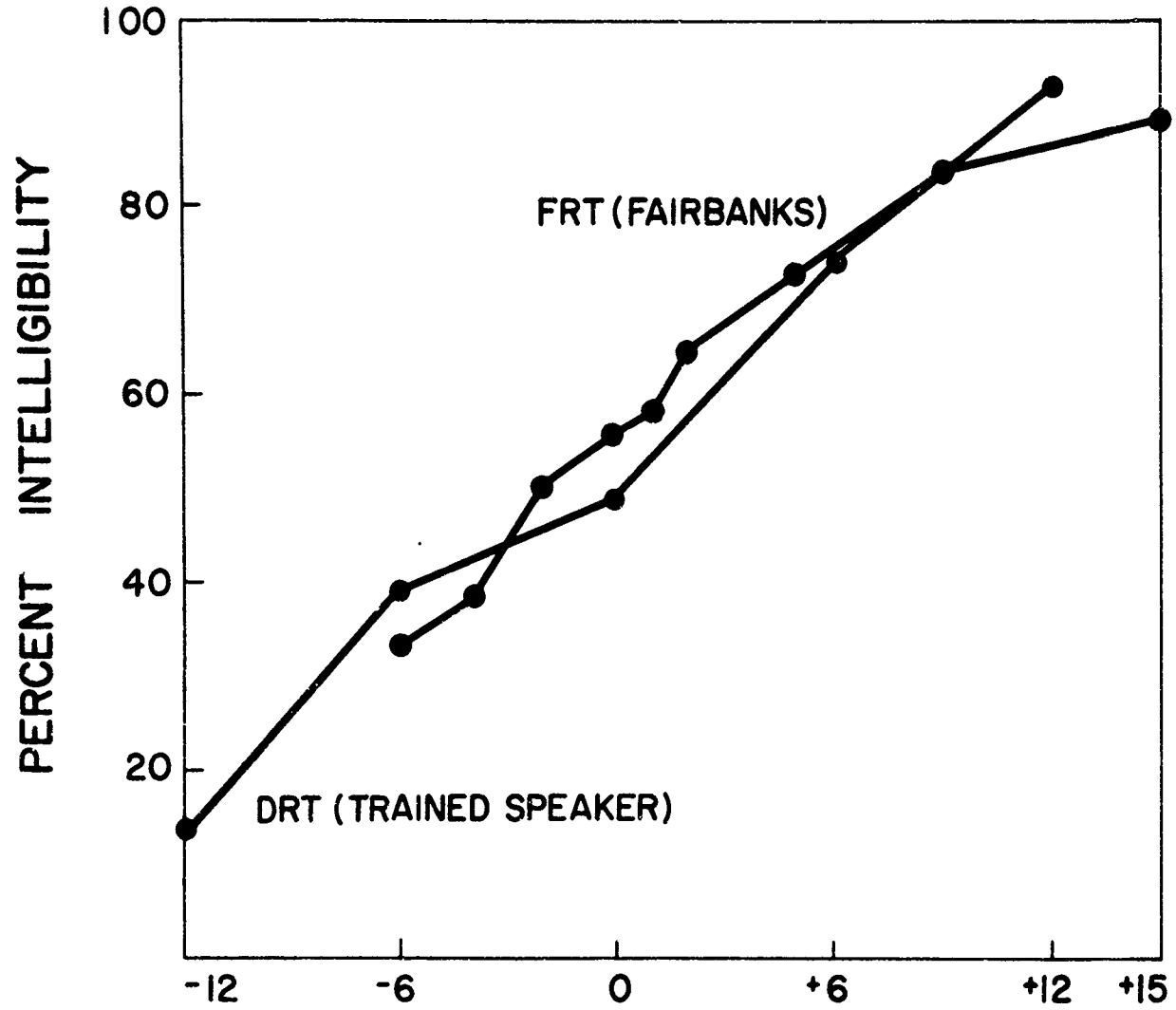


FIG. 1.1 Effects of noise upon intelligibility test scores.

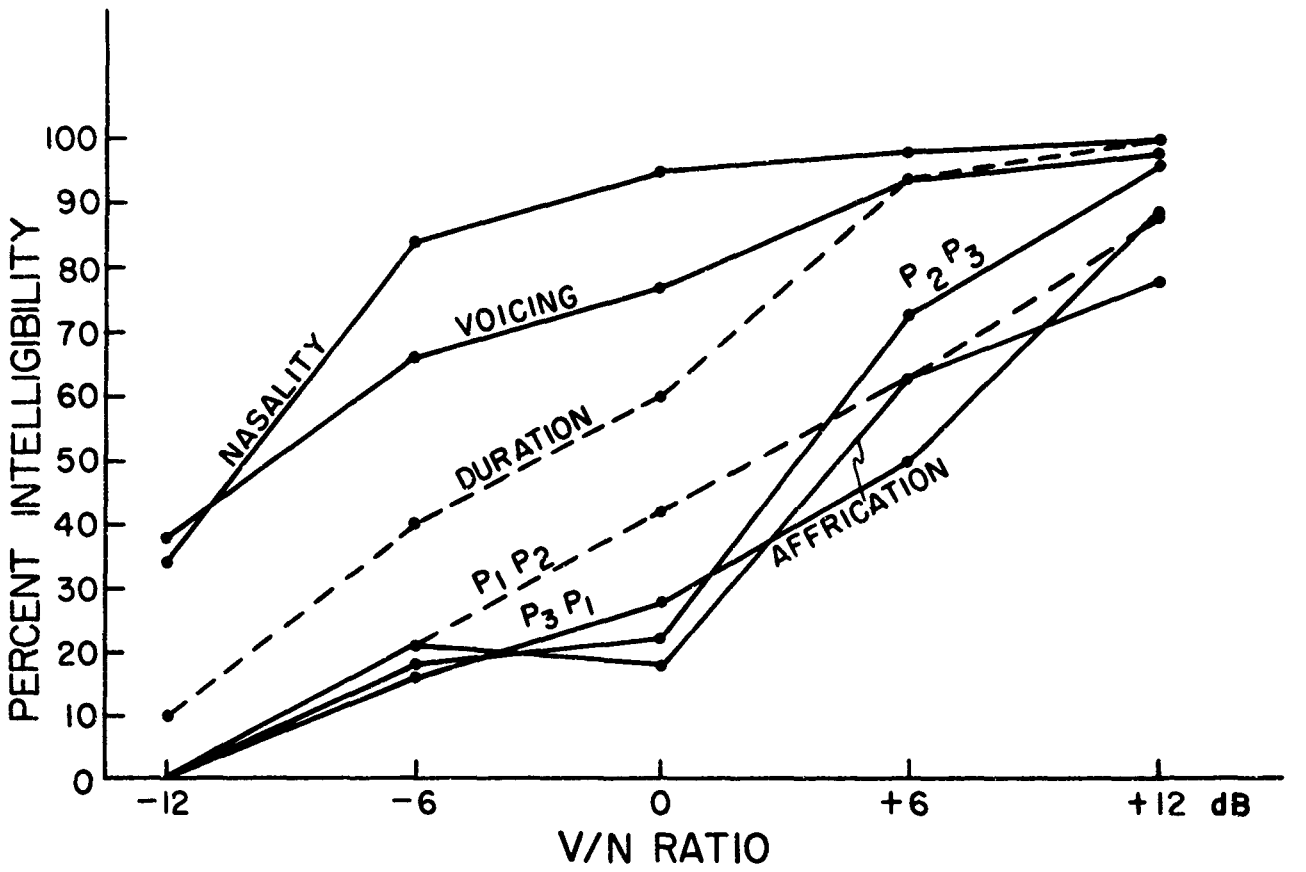


FIG. 1.2 Intelligibility as a function of vowel-to-noise ratio with consonant attribute as a parameter.

seen, white noise seriously degrades the transmission of cues leading to judgement of place of articulation and of affrication, and yet nasality is little affected. The independent behavior of the seven subtests is discussed more fully in the following section.

Figure 1.3 is a graph showing the standard error of the mean as a function of percent intelligibility. Each point represents eight subjects and 448 items. Also shown on this graph are standard errors of the mean based on results of a series of vocoder evaluations. These errors are based on eight subjects but on only 224 items and so are somewhat larger. However, the same trend to smaller standard errors is exhibited as the intelligibility scores increase.

Validity

In an effort to obtain a more sound comparison between the DRT and the Fairbanks Rhyme Test (FRT), a second experiment was performed in which both rhyme tests were presented under identical speech-to-noise conditions. For this experiment, the recordings of the random form of the DRT and the complete FRT were used. Thirty-two subjects served in this experiment (four groups of eight), and four speech-to-noise conditions were used (-9 dB, 0 dB, +9 dB, and +18 dB). Conditions were the same as those described above except that stimulus materials were filtered from 200 Hz to 4000 Hz. Each group listened once to the DRT and once to the FRT under each of the four conditions (the design is explained more fully in the Appendix).

The results of this experiment, shown in Fig. 1.4 , indicate that the FRT yields a somewhat higher score than the DRT when both are degraded by identical noise. Also apparent from this graph is the fact that in the critical range, from 60% to 100% the FRT shows a gain function of close to 2% per decibel

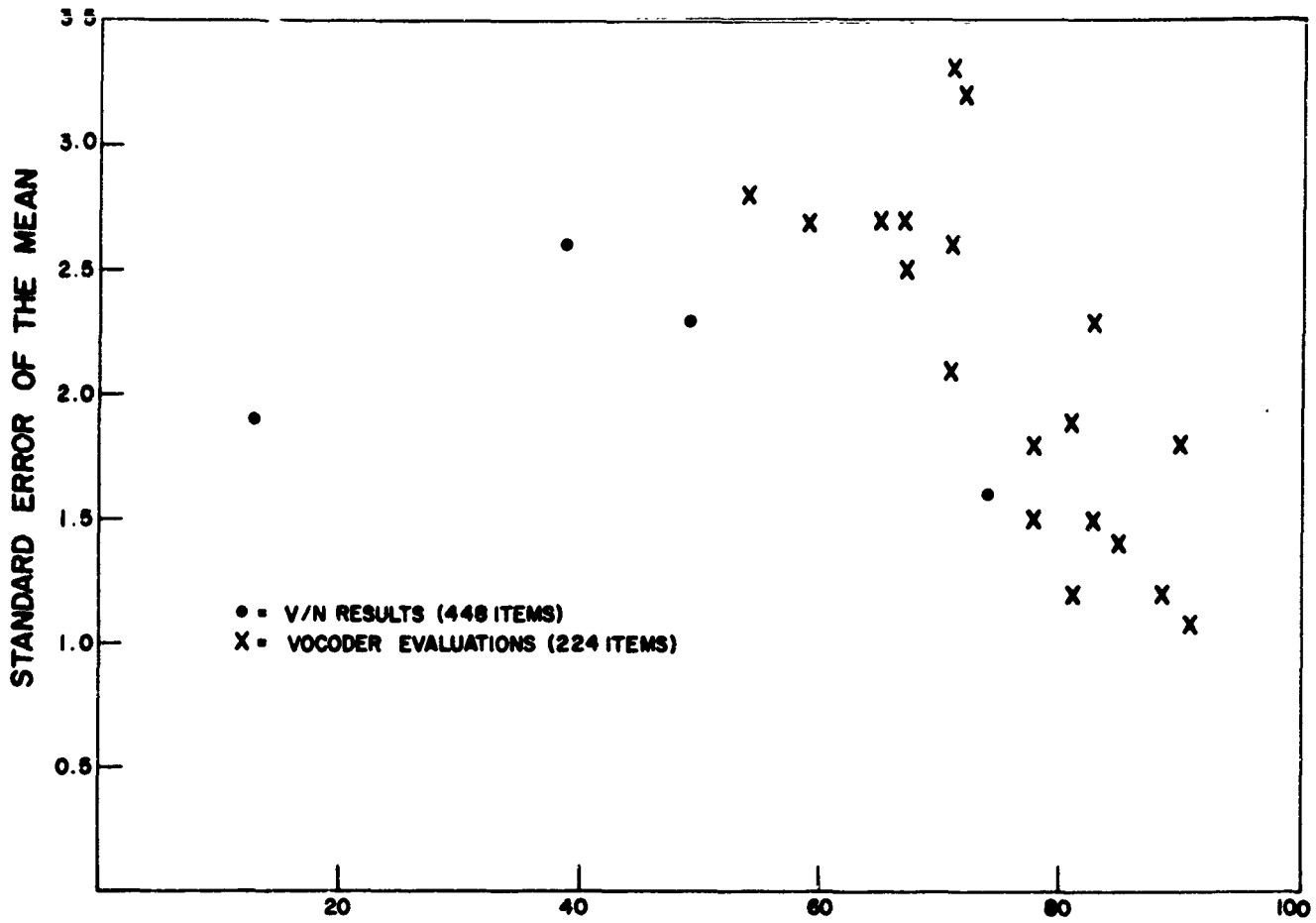


FIG. 1.3 Standard error as a function of intelligibility level.

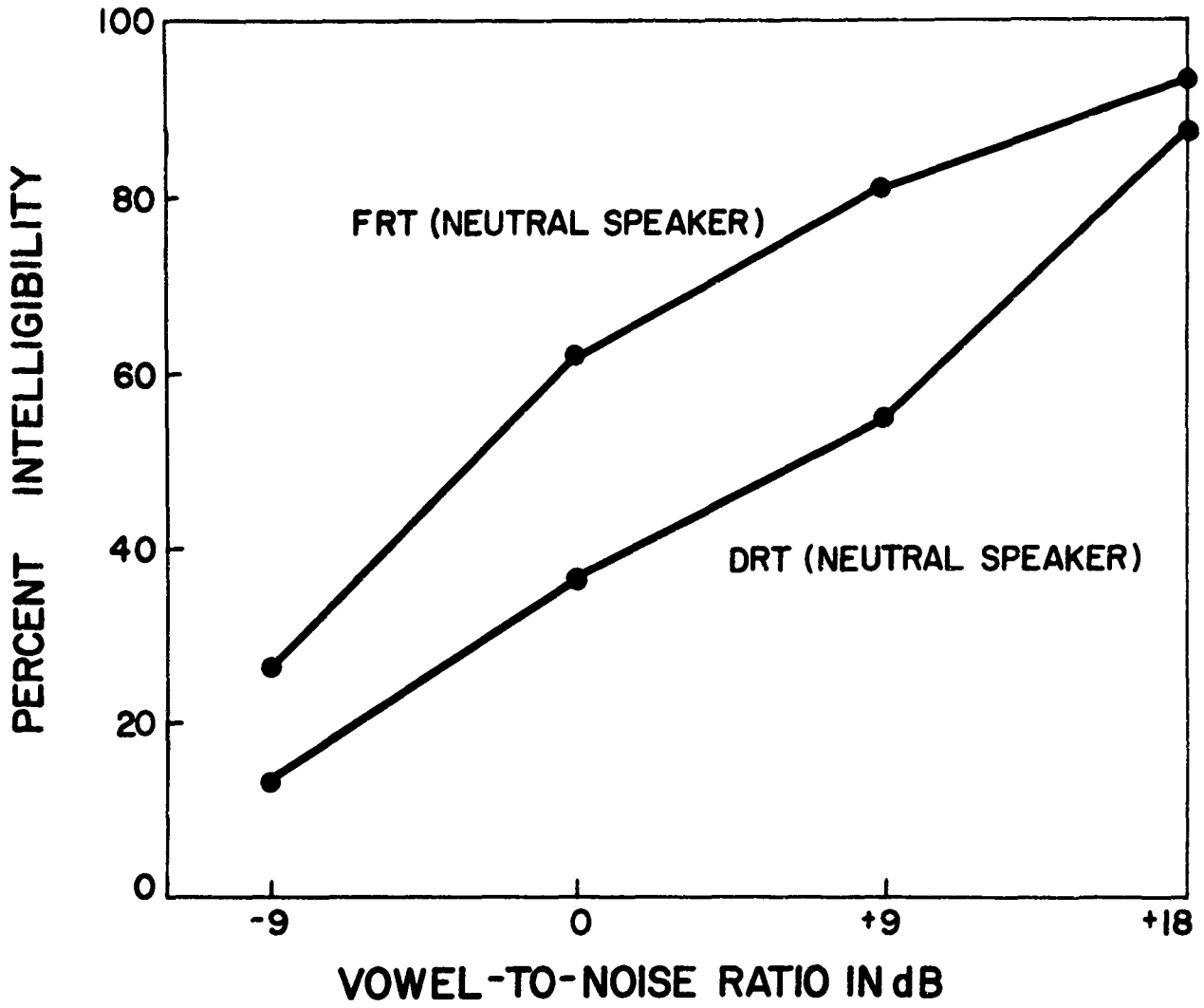


FIG. 1.4 A comparison of Diagnostic Rhyme Test and Fairbanks Rhyme Test scores under various noise conditions.

change in noise level, while the DRT shows a gain function of approximately 4% per decibel over the same range.

A further example of the relation between the two tests is provided by a scattergram showing DRT scores versus FRT scores for identical vocoders (Fig. 1.5). As can be observed for this sample of vocoders, the FRT scores extend over a smaller range (≈67% - 95%) than DRT scores (≈55% - 90%) and tend to run somewhat higher. It appears therefore that the DRT can provide a valid measure of over-all system performance.

However, the primary purpose of the DRT is to provide independent evaluations of various independent aspects of system performance.

Some demonstrations of its validity in this respect are provided by the results of several system evaluations. The results of a study in which the DRT was used to evaluate an 18-channel analog vocoder operating as a conventional vocoder, a monotone vocoder, and as a whispering vocoder are of particular interest in this connection. These results are presented in Fig. 1.6. As should be expected, the three modes of this vocoder are very nearly alike on both the attribute present and absent with respect to all features, except voicing. The whisper vocoder yields a score of 72% for transmission of voicing cues, while the monotone and conventional modes yield scores of 88% and 90%, respectively.

A study of the effects of multiple vocoderization upon speech intelligibility provides another example of the experimental independence of various diagnostic scores. Recordings of DRT materials as processed by each of three experimental vocoders were used in this study. Initial output

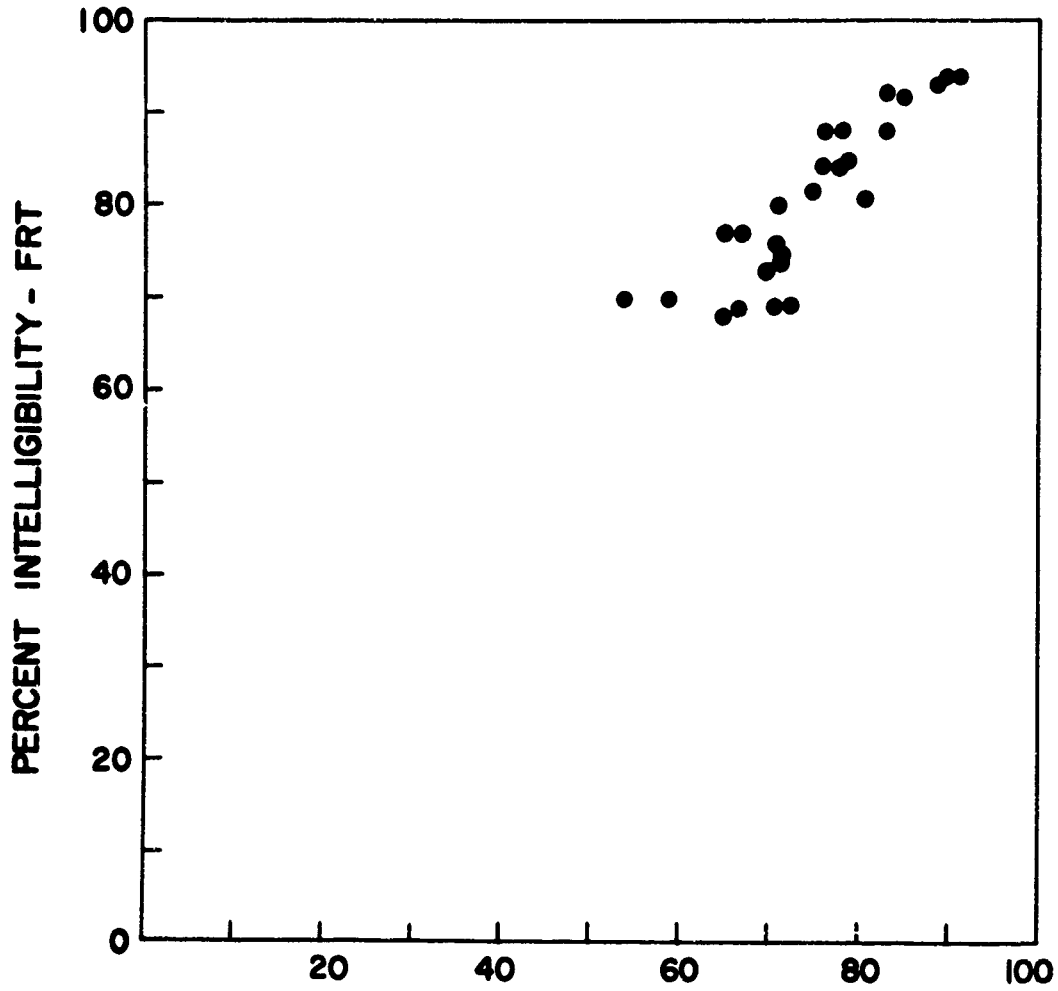


FIG. 1.5 Scattergram of Fairbanks Rhyme Test scores vs Diagnostic Rhyme Test scores for a sample of channel vocoders.

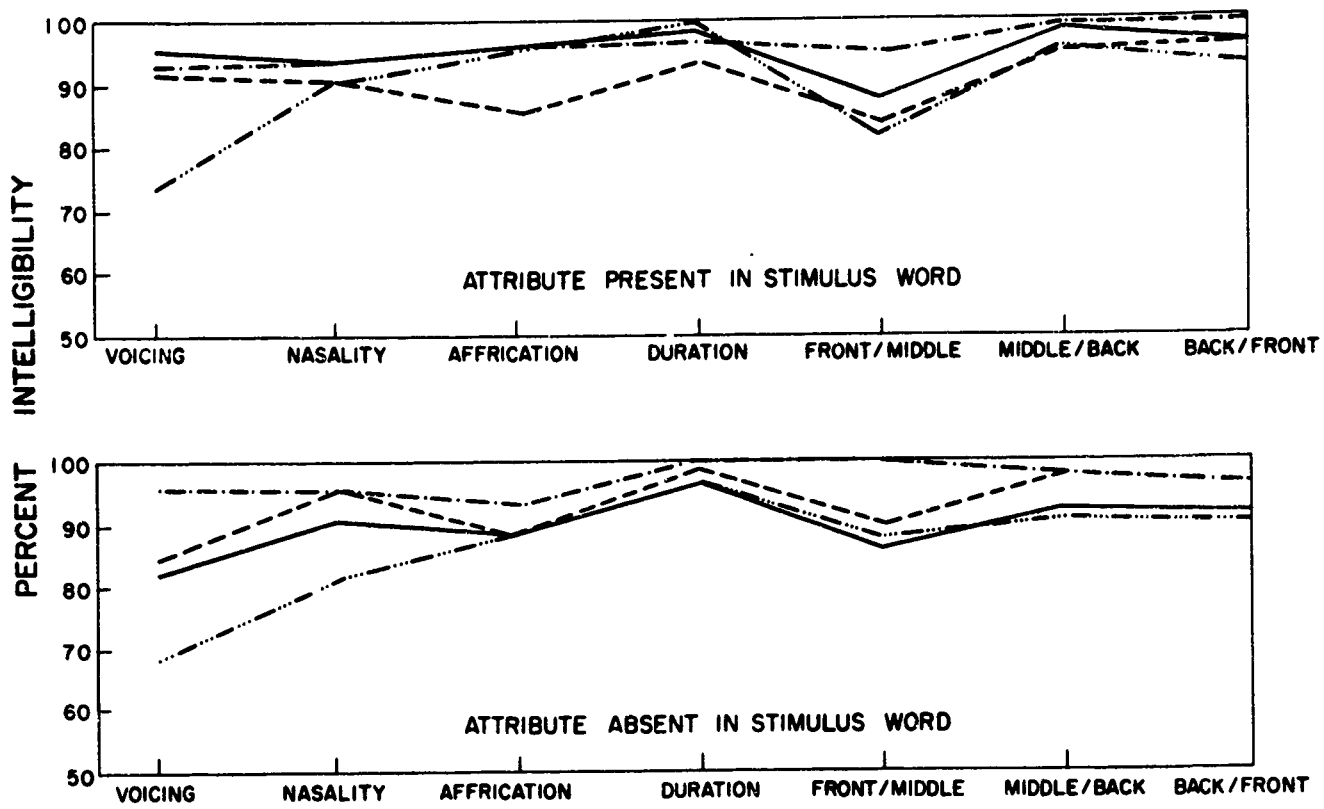


FIG. 1.6 Diagnostic scores for three types of vocoderization.

recordings were used as inputs and processed again by the same vocoders. This was done one more time to obtain recordings of the DRT as vocoderized three successive times. The results of an evaluation of these tapes are shown in Fig. 1.7. Each graph, representing one vocoder, shows intelligibility scores as a function of the number of vocoder "passes" with consonant attribute as a parameter.

In general, the trends revealed here are consistent with expectations. Scores on all of the diagnostic dimensions tend to decrease with successive degrees of vocoderization though the rate of decrease varies with the attribute as well as with the vocoder involved. There are, moreover, some exceptions to these general trends, with the possible implication that tests of multiply-vocoded speech can reveal system deficiencies for which conventional testing procedures are insensitive.

Speaker Effects

Just as intelligibility scores of a given speech transmission system will vary with the particular test used to obtain the score, so also may scores be expected to vary when different speakers are used for identical test materials. In particular, one might expect speaker differences to find expression in the various diagnostic scores of the DRT. Accordingly, an experiment was performed to determine the effects of speaker differences on DRT scores. Six speakers were selected to represent the extremes of three perceived voice characteristics (see Chapter 3). These three characteristics are: pitch-magnitude, loudness-roughness, and clarity-beauty. Also a speaker, judged by a listening crew to be neutral, and a trained speaker were used. Each of the eight speakers recorded one DRT list. The recorded materials

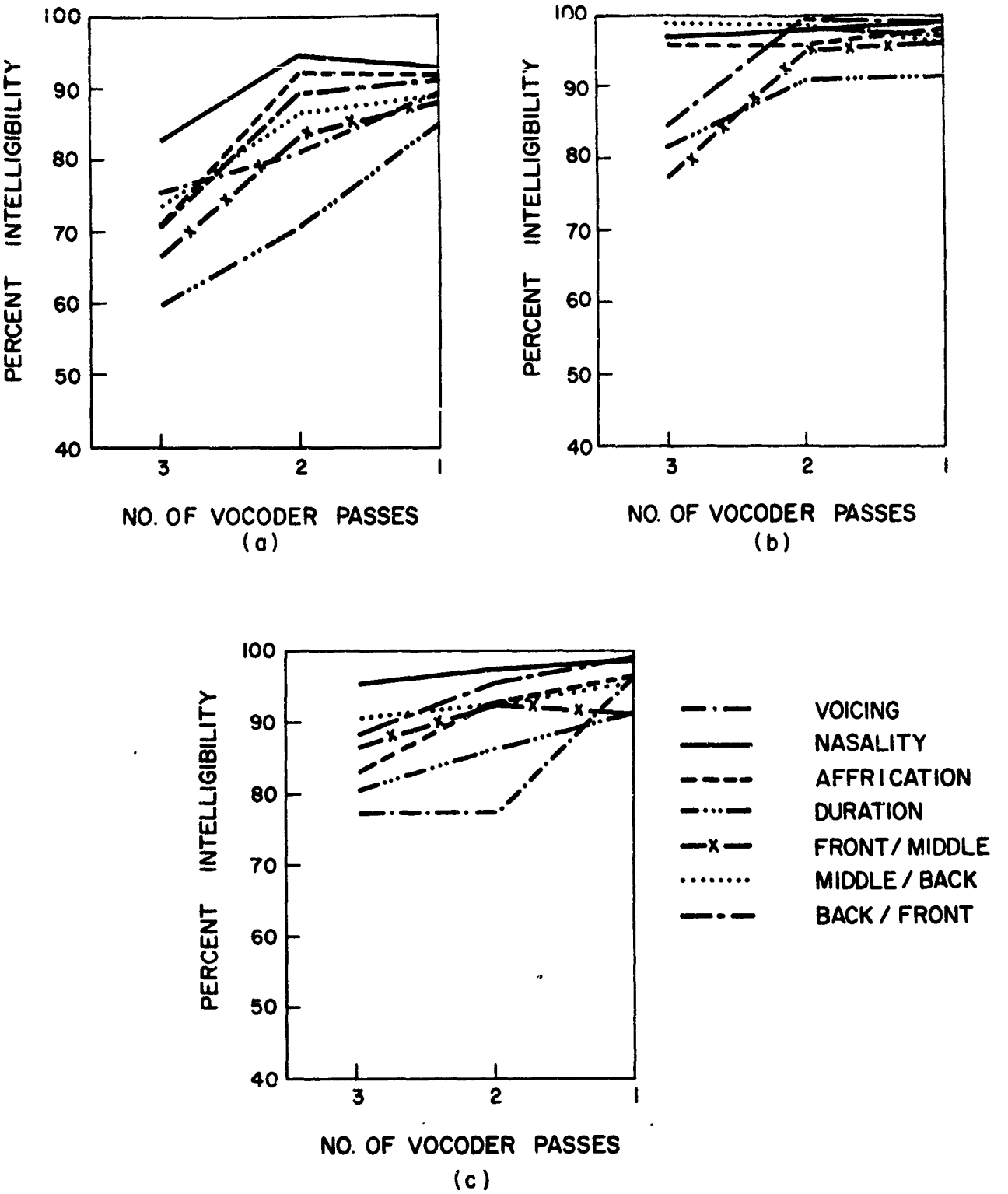


FIG. 1.7 Effects of multiple vocoderization upon Diagnostic Rhyme Test scores.

(a) experimental pitch-excited vocoder

(b) experimental voice-excited vocoder

(c) experimental pitch-excited vocoder

thus obtained were processed by four experimental analog vocoders, and output recordings for each vocoder were presented to a crew of eight listeners.

Diagnostic scores for eight speakers, as averaged for four vocoders, are presented in Table 1.2. As can be observed, there is variation among the intelligibility scores yielded by the eight speakers, and this variation appears to be somewhat predictable. The "trained" speaker and the "clear" speaker are quite consistent in yielding the highest scores, while the "rough" speaker yields a relatively low score.

A coefficient of concordance was computed in order to evaluate, more generally, the consistency of inter-speaker differences across vocoders. The obtained value of 0.65, $P \leq 0.13$, is strongly suggestive of differences in the inherent intelligibility of individual voices. This value is not sufficiently high, however, to rule out a significant interaction of vocoders with voices. Independent of their inherent differences in intelligibility, some voices may be better adapted than others to the transmission characteristics of individual vocoders. It would seem that this is particularly true in regard to the individual diagnostic features. It is apparent from cursory examination of Table 1.2 that the Diagnostic Rhyme Test is sensitive to individual differences in "vocodability." The most obvious example of this occurs for voicing, where Speaker 6 (loud-rough) exhibits a relatively low score. For affrication, however, this same speaker exhibits relatively high scores. The results of this experiment clearly indicate a need for further investigation of speaker effects on intelligibility scores.

The DRT has been used to evaluate a number of vocoders representing a fairly broad range of data rates as well as a diversity of speech synthesizing techniques. The results of these evaluations provide additional evidence of the general sensitivity of the DRT to qualitative and quantitative

Table 1.2 Diagnostic Scores For Eight Selected Speakers (Averaged Over Four Vocoders)

Speaker	Voicing	Nasality	Affrication	Duration	Front/Mid	Mid/Back	Back/Front	Mean
1. Trained	97	96	90	99	38	98	98	95
2. Neutral	95	93	94	90	92	97	96	94
3. Low	81	93	92	95	83	94	95	91
4. High	94	94	91	94	86	87	96	91
5. Smooth	96	96	88	96	86	94	90	92
6. Rough	85	94	94	91	78	88	90	88
7. Clear	95	95	92	93	89	96	95	94
8. Unclear	93	98	93	92	88	95	95	93

Note: $\sigma_e^2 \cong 2.0$ for individual diagnostic scores. $\sigma_p^2 \cong 0.5$ for means.

differences among speech processing devices. They also provide valuable insights into the characteristic deficiencies of various vocoders.

Consider first the qualitative and quantitative implications of digitalization as revealed by a comparison of results from Figs. 1.8 and 1.9(b). While not representing the analog and digital versions, respectively, of a single set of vocoders, the two figures represent more or less equivalent samples from the same vocoder pool. Digitalization is thus the major feature distinguishing the vocoders of Fig. 1.9(b) from those of Fig. 1.8.

Of primary interest, perhaps, is the gross cost of digitalization as evaluated by a comparison of total DRT scores for the two cases of interest. It appears that the primary consequence of digitalization is typically a loss of the order of 15-20 percentage points in over-all intelligibility. While some of this loss can be offset by an increased bit rate in conjunction with the use of voice excitation, the gains would not seem commensurate with the price. A fourfold increase in bit rate would, at least in this instance, seem to require justification on grounds other than increased intelligibility. On the other hand, reduction from a data rate of 2400 bps to 1200 bps would appear to have relatively minor consequences for intelligibility. However, this latter conclusion should be accorded only the most tentative acceptance because of the small number of 1200 bit vocoders involved here.

An examination of the diagnostic score patterns of Figs. 1.8 and 1.9(b) reveals that the effects of digitalization are not equally severe for all consonant attributes.

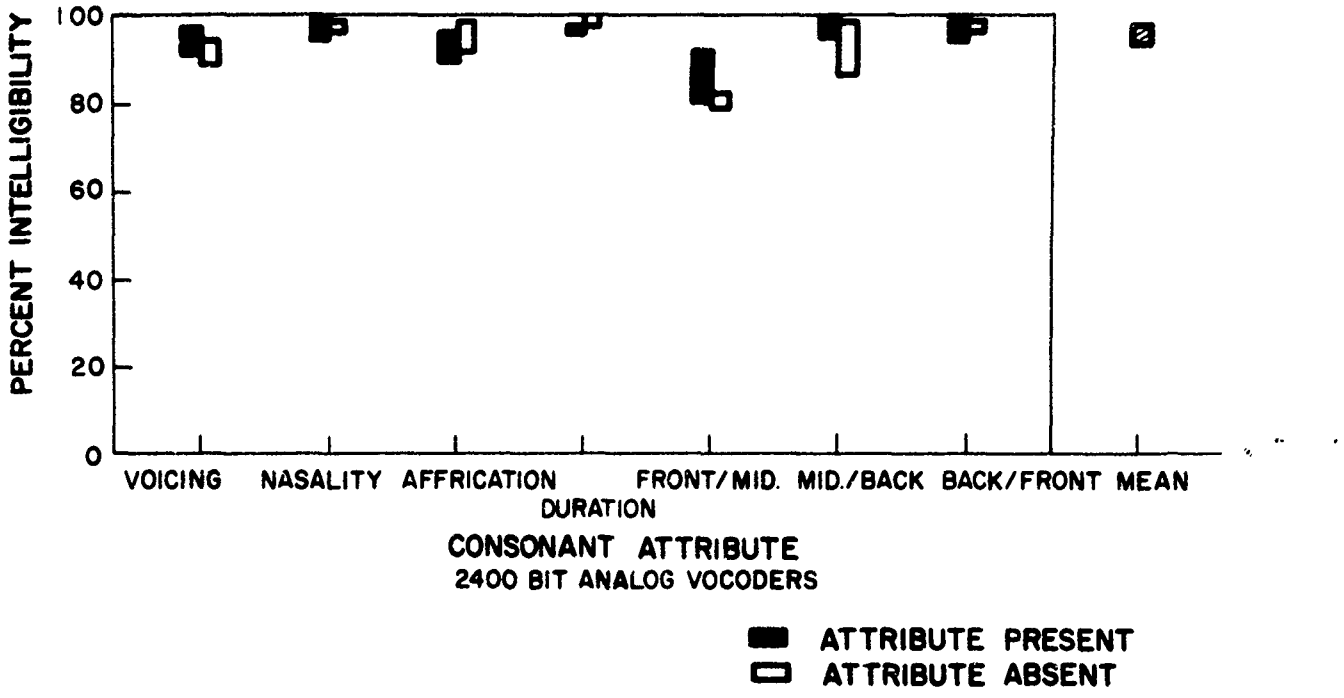


FIG. 1.8 Ranges of diagnostic scores for a selection of 18-channel analog vocoders.

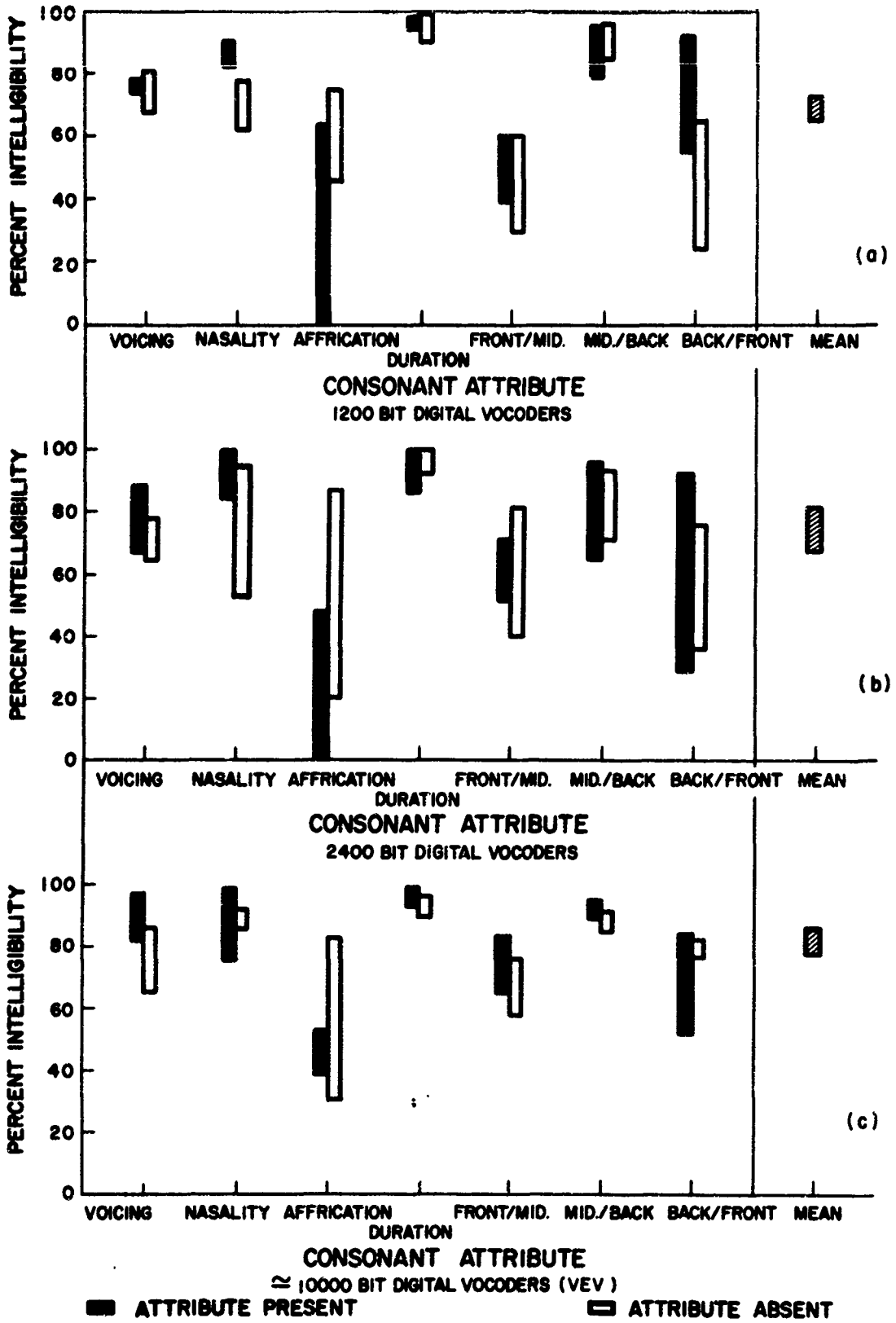


FIG. 1.9 Ranges of diagnostic scores for selected digital vocoders.

Duration appears to be relatively unaffected by any of the various forms of speech processing involved here, although it is not insensitive to other forms of speech degradation, or, for that matter, to all forms of vocoderization. The transmission of all other attributes is significantly affected, though in varying degrees, by digitalization.

Voicing information is lost in the course of digitalization, although it would appear that these losses can be compensated for in some degree by increased data rate, particularly in conjunction with voice excitation.

Information with respect to the attribute, nasality, is lost through digitalization. While it would appear that this loss can also be compensated for to some extent, a tendency to "overnasalize" seems to be an inherent characteristic of digital vocoders.

Fricative information is clearly the most vulnerable to degradation by digitalization, particularly so with regard to information concerning the presence of affrication. Among the attributes relating to place of articulation front (vs middle) and back (vs front) appear to sustain greatest degradation as a result of digitalization. Only the former tends to improve consistently with increasing data rate.

From the results in Figs. 1.3 and 1.9, it would appear that the discriminability of the "place attributes", middle (vs back) and back (vs front) are effected differently by the various types of speech processing represented in the figures. Thus, in spite of their equivalence (except for polarity) in the Jakobson-Halle taxonomy, they would appear to represent experimentally independent parameters of vocoder performance. However, further research on this issue-and more generally the issue of the experimental independence of the various DRT scores - is undoubtedly warranted and will be undertaken when a sufficient amount of vocoder test data has been accumulated.

Summary and Recommendations

The results presented above attest to the inherent reliability and validity of the Diagnostic Rhyme Test and of the various scores which it yields. In the course of the research conducted thus far, each of the 14 subscores pertaining to a consonant attribute has served on one occasion or another to reveal a design deficiency or malfunction in an experimental vocoder. Moreover, the validity of the DRT total score, as a measure of gross system performance, has been established empirically. Thus, the Diagnostic Rhyme Test, at its present stage of development, has proved more than adequate to serve the purposes for which it was designed.

There exist , however, a number of possibilities for the further refinement of the Diagnostic Rhyme Test as well as for the extension of its range of applicability. While the results obtained thus far suggest a number of correlations between various diagnostic scores and specific vocoder design features, types of malfunction, etc., the nature and degree of these various correlations remain to be firmly established. Data collected in the course of routine evaluations made with the DRT will, over time, contribute significantly to this end. However, there is clearly a need for studies involving the systematic experimental manipulation of vocoder parameters. Minor modification , designed primarily to eliminate some of the redundancy of the present DRT materials, is to be desired at some time in the near future. However, it should only be accomplished with careful experimental checks to ensure the essential comparability of scores yielded by present and future versions of the DRT.

Further research is needed on the roles of different types of listener-experience with DRT materials. While all results obtained thus far

indicate the DRT to be generally insensitive to listener practice, questions of the effects of listener familiarity with individual speakers, specific utterances, etc., merit further attention.

The role of speaker differences also merits further attention, particularly as such differences manifest themselves on the various subtests of the DRT. The results obtained thus far suggest that ratings received by speakers on various perceived acoustic traits can be of value in predicting a speaker's inherent over-all intelligibility as well as his characteristic diagnostic score pattern. However, much further research is needed on this issue.

Clearly, the topic of consonant-vowel interactions in relation to consonant intelligibility deserves further consideration. It would appear that diagnostic scores obtained with the present combinations of vowels can be generalized to the case involving all vowels, and that we have effectively sampled the population of English vowels in a representative manner. It would also appear, however, that the recognizability of consonants generally, as well as individually, varies significantly as a function of the particular vowel involved. It is conceivable, therefore, that the effects of a given form of speech processing upon consonant intelligibility may,in some instances, be conditional upon the vowel involved. More to the point, examination of listener performance with respect to certain CV combinations may reveal otherwise undetectable system deficiencies. Future research with the DRT should thus take adequate account of such possibilities as these.

Finally, some further efforts seem warranted to adapt the principles and meterials of the DRT for use in a "free conversation" testing procedure.

CHAPTER 2

SYSTEM EVALUATION FROM THE STANDPOINT OF SPEECH QUALITY

At present there are a number of fundamentally different ways in which speech may be processed to reduce the channel space required for its transmission. The critical system parameters-sensitivity, noise level, bandwidth, etc. vary greatly. Moreover, it is generally difficult to establish systematic relations between these various parameters and what is conventionally referred to as speech "quality." This does not, however, preclude quantification of the subjectively significant characteristics of speech as processed by such systems. Various methods designed to accomplish the latter have, in fact, been employed in the past, although none has proved altogether satisfactory. Moreover, a number of fundamental issues have yet to be resolved before any completely satisfactory solution can be achieved.

One of these pertains to the meaning of the term "quality" as it is generally used in reference to the subjectively significant characteristics of speech. For present purposes no formal definition of this concept will be attempted; a pragmatic definition will suffice: Quality is defined as the totality of transmission channel characteristics which contribute to the overall acceptability of speech, or more specifically, of a particular form of speech processing. By definition, one form of processed speech possesses a higher level or degree of quality than another to the extent that the typical listeners would prefer it as a general means of voice communications.

Ultimately, it would be desirable to amend this definition of quality to include the condition "other things being equal", particularly where the "other things" are intelligibility and speaker recognizability. However, until the means for the operational realization of this more restricted conception of quality are available, - i.e., until practical methods

of partialling out the effects of intelligibility and speaker recognizability have been developed – the broader conception of quality must be retained. It is with this more global conception of quality, therefore, that this present report deals.

A second major issue concerns the dimensionality of "quality," particularly where we are dealing with "quality" in the broadest sense of the term. Specifically, questions can be raised concerning the qualitative stability of the criteria employed by listeners in judging or comparing two or more speech processing conditions on the basis of quality. Several investigators (e.g., McGee, 1961) have reported evidence in support of the possibility that the subjective bases of "preference responses" are multidimensional in nature, that the subjective basis of an individual preference response may vary from one context, if not from one instance, to the next. Such a possibility justifies some doubt concerning the psychological meaningfulness of values representing the relative "over-all preferability" of various forms of processed speech.

Regardless of theoretical significance of such questions, their practical significance hinges upon a single issue: the possibility of a unidimensional measure of over-all acceptability which, by some simple transformation or another, permits predictions to some pragmatic criterion of listener tolerance or acceptability for a given form of processed speech.

In other words, this is an issue of whether a "psychological distance function" can be found such that observed inter-system or inter-condition distances can be reconciled within a single dimension, or whether additional dimensions are generally necessary to account for "preferability differences" among two or more speech-processing systems or conditions.

In general, the absence of "transitivity" in comparative preference values is taken as evidence of multi-dimensionality. However, transitivity failure may also occur as a result of the use of an inappropriate metric for scaling preferability. Thus, lack of transitivity becomes a valid criterion of multi-dimensionality only where it can be demonstrated to hold for all possible functions of "preferability differences."

A third issue concerns the most appropriate basis for quantifying preference differences. It is thus related to the issue of dimensionality. The problem posed here is one of devising a metric for preferability which, a priori, has the greatest likelihood of satisfying the joint requirements of transitivity and unidimensionality. An examination of the possible approaches here reveals several which, though used by various investigators, (e.g., frequency averaging), can find little justification in psychophysical theory. The choices reduce ultimately to some variation of the method of pair comparisons, based upon a principle which is roughly analogous to that underlying the Thurstone "Law of Comparative Judgement." Thus, the fundamental principle on which the unit variance method rests is one whereby inter-individual differences in "preference response" distribute normally with respect to a unidimensional continuum which has the properties of an "equal interval scale." This principle only serves, however, to determine the general features of the method. A number of more specific details of the method can, at the outset, be defended on little more than intuitive grounds. All issues concerning specific details are, however, subject to experimental resolution and it was to achieve such resolution that a major part of the experimentation to be described here was designed.

Among other things, the psychophysical method of pair comparisons requires a minimum of training for the listener and provides the greatest set

of possible checks on the internal consistency of the listener's behavior.

Some amount of exploratory research was undertaken prior to the final formulation of the standard unit variance method. Initially, the test materials consisted of 28 conversational sentences which were also used as the stimulus materials in the SRRC voice recognition tests. A list of sentences is in Appendix I. A professionally trained speaker was used. Subsequently, new speech materials were introduced which consisted of 10 sentences of 8 syllables each. (An answer sheet and list of sentences are in Appendix I.) Five speakers were selected from the pool of speakers previously used in studies of speaker recognition. These included five voices which were judged as being "neutral", "low-pitched", "loud-rough" and "soft-smooth". Essentially the same procedures were employed in preparing "working" master tapes for all of the methods investigated.

All recording was done as previously described. All sentences were recorded on one-half inch magnetic tape (Ampex 600 series 631-273111) using a Crown 1400 series recorder.

From the master recording a two-channel, quarter-inch magnetic tape (Scotch 1/4-138-12 1-1/2 mil polyester) was prepared, such that one of the identical utterances of the same sentence was copied on channel A and the other on channel B. The order of channel selection for each pair of utterances was randomly determined. Regardless of the number of sentences used, channel A preceded channel B in one-half of the presentations; conversely, channel B preceded channel A in one-half of the presentations. The pattern in which utterances were distributed between channels was identical for all five speakers. Ten different orders of sentences were used, each consisting of five speakers uttering two pairs of sentences. The order in which speakers were grouped within each set of 10 sentences was randomly determined. This working-master tape thus provided the input materials used in evaluating ex-

perimental speech-processing devices. More specifically, these were used as follows: Outputs from one vocoder were equalized by adding bass or treble so that the speech materials sound most pleasing to several listeners. Then, these bass and treble settings for that vocoder were used as a standard for equalizing other vocoder materials. Speech materials from channel A were played through the equalized vocoder mode and recorded on magnetic tape. Using the reference standard, other vocoder modes were equalized for bass and treble as closely as possible.

Final preparation of test magnetic tapes consisted of pairing different vocoders. For n vocoders there were n(n-1)/2 pairs. Two Crown SS-800 series tape recorders were used to play back the vocoderized speech materials from channel A and from channel B using the desired pairing of vocoders. Loudness levels and equalization for bass and treble were adjusted before recording the speech materials onto one-half inch magnetic tape using Crown 1400 series tape recorder. Two different vocoders comprised a pair which, in turn, contained two sentences uttered by each of five speakers.

The Unit Variance Scaling Procedure

Munson and Karlin (1962) suggested the use of a variable reference system which would be degraded along some known physical parameter until it was judged equal to the unknown system. In addition, the reference system needed a wide range of physical parameters because it had to be preferred over the unknown system. The most convenient variable reference system was signal-to-noise ratio. In the early phases of this program a vowel-to-noise ratio was used as a standard. Psychological functions were obtained using the method of constant stimuli. The isopreference point was then obtained

for the unknown systems. The use of the vowel-to-noise ratio as a standard was abandoned because of the prohibitive amount of work involved in testing a large number of systems and because of the difficulty of the listener's task.

A set of fixed standards was chosen. These consisted of four "representative" digital vocoders. These four vocoders are used in all experiments in which experimental vocoders are evaluated. The usual experiment thus consists of 6 vocoders of which two are experimental. There are two advantages in using such a standard. First, all vocoders are directly compared to each other; second, the four standard vocoders provide a constant context from one evaluation to the next. At present, procedures are standardized such that scale values for each vocoder are based on 800 responses - 8 listeners x 5 speakers x 2 trials x 5 vocoder pairs x 2 sentence per speaker. Since each sentence is spoken every three seconds, the entire experiment lasts less than one-half hour.

Unit Variance Method

The Unit Variance method of scaling derives its name from the fact that the estimated true variance among listeners in their evaluation of the quality differences provides the basis for the establishment of a scale unit.

The steps required to obtain scale values using the Unit Variance method are described in detail below.

1. Convert observed frequencies into proportions for each individual speaker and for each pair of directly compared vocoders.

2. Convert proportions into arcsine values (in radians).

TABLE 2.1

Analysis of variance for two directly compared vocoders and one speaker

Entry	Sum of Squares	df	Mean Square
Rows(Listeners)*	$SSR = \sum_{i=1}^{R}\left(\sum_{j=1}^{C}\Sigma x^2\right)\Big/C - T^2/N$	R-1	SSR/R-1
Columns(Trials)*	$SSC = \sum_{j=1}^{C}\left(\sum_{i=1}^{R}\Sigma x^2\right)\Big/R - T^2/N$	C-1	SSC/C-1
Residual	SSRC = SST - SSR - SSC	(N-1)-[(R-1)+(C-1)]	$\frac{SSRC}{(N-1)-[(R-1)+(C-1)]}$
Total	$SST = \sum_{i=1}^{R}\sum_{j=1}^{C}\Sigma x^2 - \left(\sum_{i=1}^{R}\sum_{j=1}^{C}\Sigma x\right)^2\Big/N$	N-1	

* Rows = 8 listeners
Columns = 2 trials

3. Use analyses of variance for each of the five speakers and for each pair of directly compared vocoders to obtain true variance among listeners.

4. Obtain an unbiased Z score between two directly compared vocoders and for each speaker individually.

5. Obtain a mean Z score for each speaker using those pairs of vocoders in which one vocoder is common.

6. Obtain a mean Z score of five speakers for a given vocoder.

7. Divide the obtained scale values by two.

To determine the scale unit, the arcsine transformation data are subjected to analysis of variance as shown in Table 2.1.

SSR/(R-1) has a sampling distribution with mean equal to $\sigma^2 + n\sigma_m^2$, where $\sigma_m^2 = \sum_{i=1}^{k} (\mu_i - \bar{\mu})^2/(k-1)$. The residual variance SSRC/{(N-1)-[(R-1)+(C-1)]} is an unbiased estimate of σ^2 . The value of σ_m^2 may be estimated by subtracting the residual variance from the mean square for means and dividing it by n . Thus SSR/(R-1)-SSRC/{(N-1)-[(R-1)+(C-1)]}/n is an unbiased estimate of σ_m^2 (Dixon and Massey, 9). The distance between two vocoders is for a theoretical listener and an infinite number of trials.

The scale value for any one vocoder is obtained from "unbiased" Z-difference scores of several pairs of vocoders. Specifically, the scale value for a given vocoder A is given by:

$$Z_A = \frac{Z_{AB} + Z_{AC} + Z_{AE}}{4}$$

The scale value for vocoder B would be

$$Z_B = \frac{Z_{BA} + Z_{BC} + Z_{BE}}{4} \quad \text{etc.}$$

The separation between any two directly compared vocoders is obtained for each of the five speakers individually. There is a choice between two procedures at this point: the first involves calculation of scale values for each vocoder and speaker; the second involves pooling the data for the different speakers to obtain a single, typical scale value for each vocoder. Where speaker effects are not of concern, the second procedure is used.

Standard Unit Variance Scale (STUVS)

The Standard Unit Variance Scale is a variation of the basic Unit Variance Scale designed to provide maximum precision in the prediction of relative preference frequencies for any two vocoders or other systems which have been scaled by means of the Unit Variance method. It is obtained by a simple linear transformation of the Unit Variance Scale by an experimentally determined factor (determined on the basis of observed scale values for the four standard vocoders).

Table 2.2 shows the Standard Unit Variance Scale (STUVS) for four standard vocoders and two experimental vocoders. The STUVS is based on the mean obtained from three independent experiments. The standard deviations are based on adjusted scale values which were obtained in eight experiments where the standard four vocoders were included.

The entries in column No. 1 are STUVS for four vocoders (A,C, L, and F) and their standard deviations.

These scale values represent points on the Gaussian distribution where the scale unit is based on the true standard deviation of scores, representing measures of an individual preference for a given condition or vocoder.

TABLE 2.2

The Standard Unit Variance Scale (STUVS) -

its relationship to the observed scores[x]

Standard Vocoders	1 STUVS		2 Difference scores for STUVS	3 Percent	4 Observed Difference Scores	5 Percent
A	.7122 ± .1304	$\overline{A-C}/2$	.4154	.66	.3319	.63
C	-.1185 ± .2291	$\overline{A-L}/2$	.5030	.69	4677	68
L	-.2938 ± .1625	$\overline{A-F}/2$	.7703	.78	.7388	77
F	-.8284 ± .1304	$\overline{C-L}/2$	.0876	.53	-.0753	.47
		$\overline{C-F}/2$	.3550	.64	.3319	.63
		$\overline{L-F}/2$	.2673	.60	.2533	.60
B[xx]	.5100					
E[xx]	.0186					

x Explanations of each entry appear in the text.

xx These two vocoders are replaced by the unknown vocoders which are to be tested

Aside from the theoretical advantages of the procedures employed to derive STUVS for a set of vocoders, there is the practical advantage that this scaling procedure yields precise estimates of the percent of listeners who will favor any one vocoder over any other vocoder. Thus, to estimate the relative preferability, i.e., of listeners preferring vocoder A over vocoder B, three simple steps are required:

1. Subtract lower scale value from the higher.
2. Divide this difference by two and enter $\frac{x}{\sigma}$ column of a table of normal curve areas to obtain the "percent of cases" falling between $\frac{x}{\sigma}$ and the mean of the Gaussian distribution variable.
3. Add 0.50 to the above value to obtain the estimated percent of listeners who will prefer vocoder A over vocoder B.

Numbers in column No. 2 of Table 2.2 give the difference in vocoder scale values using STUVS entries (A-C/2, A-L/2, etc.). These differences in scale values represent the percent of cases falling between $\frac{x}{\sigma}$ and the mean (mean = 0). Converting these differences into percentages, it is possible to predict what percent of listeners would prefer one vocoder over the other. It must be stressed, however, that these differences are based on STUVS and are, therefore, predictions based on indirect calculations.

Numbers in column No. 3 are the predicted percent of listeners who would prefer one vocoder over the other.

Numbers in column No. 4 of this table are observed Z difference scores obtained from direct pair comparison of the same vocoders as those in column 2. These Z difference scores for each pair of directly compared

vocoders were included. Numbers in column No. 5 show the percent of listeners who preferred one vocoder over the other when these two vocoders were directly compared.

Figure 2.1 shows the relationship between Z difference scores obtained by direct comparison of two vocoders and the predicted Z difference scores obtained from STUVS.

Figure 2.2 shows the STUVS and standard deviations of four standard vocoders. The standard deviations for each vocoder were calculated using adjusted scale values from eight experiments in which the four standard vocoders were included.

Thus, the STUVS have the desired properties of a unidimensional psychological distance function. The data obtained from STUVS suggest that the basic requirement of transitivity, essential for unidimensionality, is satisfied. In addition, the correlation ($r = 0.9851$) between STUVS and the observed data suggests unidimensionality very strongly. Aside from the unidimensionality characteristics, the STUVS can be used to predict, and with considerable accuracy, the distribution of listener preferences for any pair of scaled vocoders.

Summary and Recommendations

The Standard Unit Variance Method for evaluating speech quality has fulfilled the purposes for which it was designed. Soundly based in psychophysical theory, it is also adapted to practical purposes of system evaluation on an outline basis. This is not, however, to deny the possibility of further improvements in the technology of speech quality evaluation. The use of "frequency of preference", as the basic datum for purposes of psychological scaling, represents an inherently inefficient use of the information contained

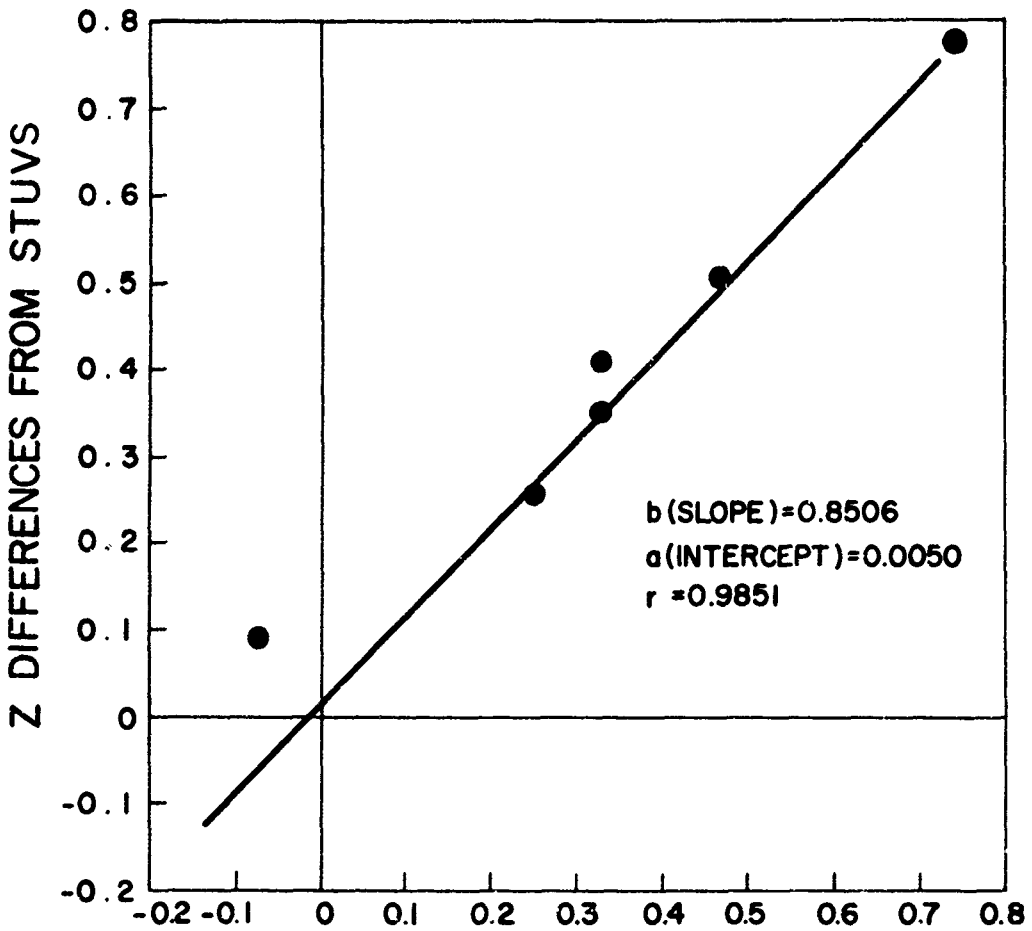


FIG. 2.1 Relationship between direct comparison distances and predicted distances.

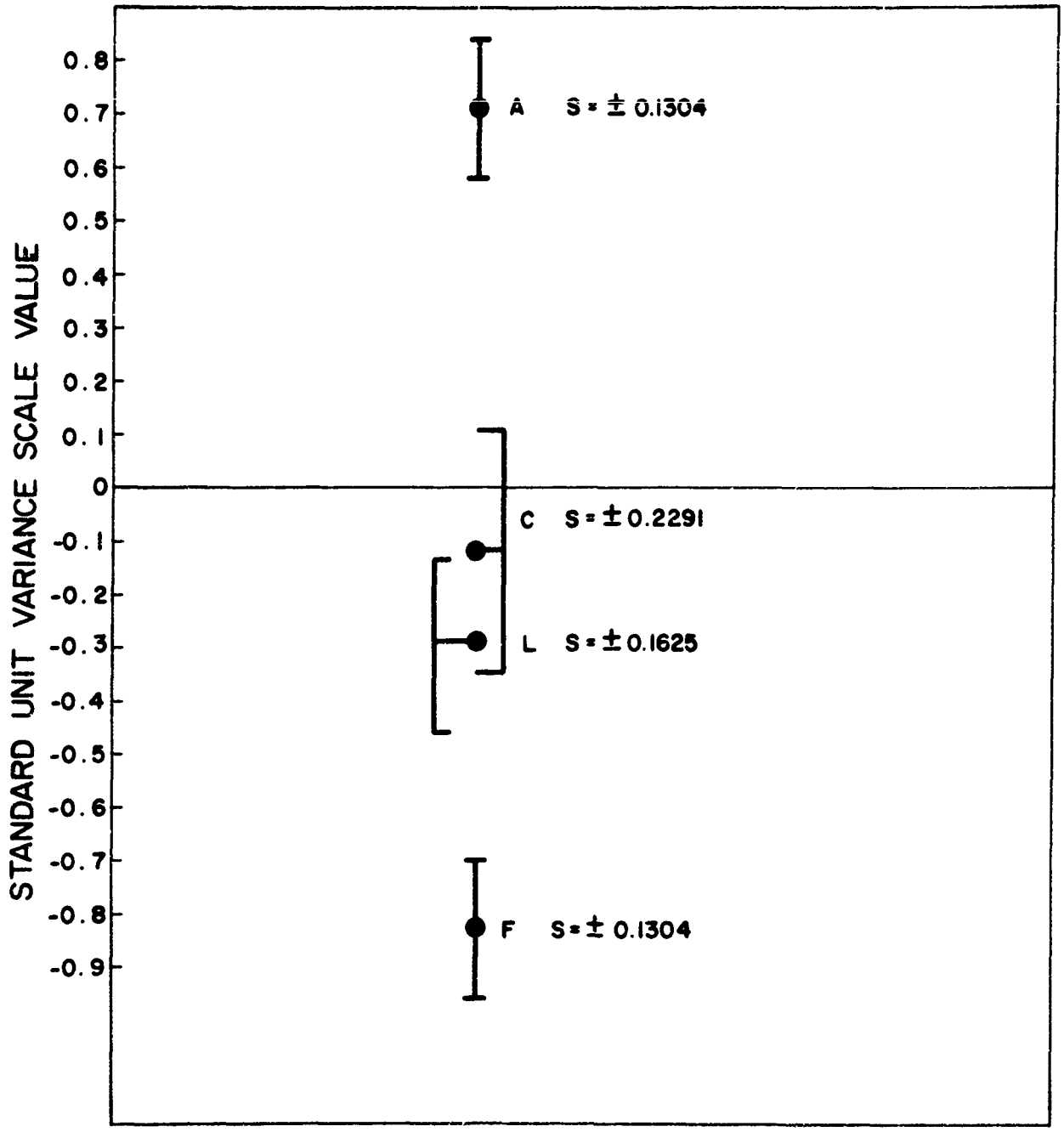


FIG. 2.2 Standard unit variance scale and standard deviations for four standard vocoders.

in a single listener's response. It is defensible on other grounds such as its superior reliability and freedom from the systematic errors inherent in the various alternative procedures. However, the use of relative ratings rather than simple preference judgements would, in general, permit the more efficient use of listener's time, given adequate techniques for stabilizing the scale of the listener's response. Some investigative effort to the end of developing such techniques would appear to be in order at this point. The STUV Method would provide an important means of validating such techniques.

CHAPTER 3

SYSTEM EVALUATION FROM THE STANDPOINT OF SPEAKER RECOGNIZABILITY

At present there is no generally accepted method of system evaluation from the standpoint of speaker recognizability. A number of different methods have been employed in the past (e.g., Shearme and Holmes, 1959, Kurtzburg, et al., 1963) while still others have potentialities which have yet to be evaluated. Although some of these methods have unique advantages to recommend them, none is without serious limitations from the standpoint of theoretical soundness or from the standpoint of practical adaptability to the circumstances in which the development of a modern communication system takes place. It may be worthwhile, however, to examine the most commonly employed method for illustration of some of the more important of these theoretical and practical issues. This method is one in which the listener's task in the operational situation is simulated more or less realistically in the laboratory. In one variation of the method, the listener is simply required to identify samples of experimentally-processed speech of bona fide acquaintances. In another variation the listener may be required to identify previously unfamiliar voices by names which he has learned to associate with them for the specific purposes of the experiment. The most conspicuous advantage of this general approach - an element of realism - is overbalanced, however, by a host of theoretical and practical limitations.

As typically employed in the past, the "realistic" method has necessarily yielded results of highly questionable generality with respect to the population of speakers and/or listeners involved. In the absence of any basis for stratified sampling, speaker sampling has been at best only nominally random. In fact, it has tended most often to be quite arbitrary. In few

instances have speaker samples been large enough to warrant any degree of confidence in their representativeness with respect to any conceivable population of interest. Listener-sampling procedures, while perhaps not so crucial, have in general been equally arbitrary.

Given a system for classification of voices, a somewhat more adequate sampling procedure might be employed with some variation in the "realistic" method. Moreover, the use of classified stimulus voices could permit the development of various types of diagnostic evaluation procedures in addition to procedures for the gross evaluation of speaker recognizability. However, even when provided with a firmer theoretical foundation, the "realistic" approach retains a number of practical limitations.

Any method which requires the use of stimulus voices previously familiar to the listener must overcome extremely formidable problems with regard to the control of extra-stimulus factors in listener response. Among the more important of these are degree of familiarity and a priori expectations on the part of the listener. Adequate control of these variables becomes crucial, particularly where it is impractical to prepare new versions of the basic test materials for each evaluation. In the course of learning to associate names with the distinguishing acoustical characteristics of voices, listeners may also learn to associate them with other characteristics, such as their temporal positions in a series of voice samples. These, and other similar effects, may thus seriously contaminate the results of experiments which require a conventional "recognition response."

Finally, there is, in fact, some basis for questioning the validity of conventional recognition scores as criteria of system performance, particularly where the listener's task is simply to associate samples of processed speech

with their previously-experienced, unprocessed counterparts. Except where explicit provision is made to test for it, the potential discriminability of processed voices, independently of their similarity to their unprocessed counterparts, may be incompletely evaluated. Moreover, prolonged training of listeners on the system under test would be required to evaluate this potential by means of the realistic approach. The practical limitations of such a procedure for routine purposes of evaluation are perhaps evident.

An alternative to the realistic approach suggests itself when we examine the phenomenon of recognition in terms of the more fundamental processes which it involves. In simplest terms, the recognition of an object or event presupposes the discrimination or evaluation of one or more definitive attributes or characters. On the basis of such evaluation the object or event is categorized, according as the configuration of perceived attributes coincides with that of some previously-experienced object or event. However, the applicability of the concept of recognition does not contingе upon the particular taxonomy or basis of categorization employed in a given situation. While the most familiar case is one in which there exists an identity category for each recognizable object or event, (e.g., the voice of Lyndon B. Johnson), we also apply the concept of recognition to situations involving other types and degrees of categorization. Thus, we may speak of recognizing a speaker when we effect the categorization: "educated," "middle-aged," "British," "male," etc. on the basis of a sample of his speech. On occasion we even use the concept of recognition in reference to a simple binary attribute. We commonly speak of recognizing a voice as being that of a man or of a woman. Without undue violence to its meaning the concept of recognition can thus be applied to a vast range of situations involving human taxonomic or categorizing behavior, and in relation to any one of many systems for classifying a given universe of objects or events.

In all instances of a taxonomy for which the concept of recognition is valid, however, there is the implication of a transformation relating the taxonomy in question to a more fundamental or intrinsic taxonomy -- a taxonomy based on the "elementary parameters" of the universe of objects or events in question -- more specifically, upon the parameters of a particular manifestation (e.g., voice sound) of the objects or events comprising a given class or population.

Normally, therefore, correct recognition of a speaker by name, sex, age, or by other taxonomic category thus presupposes the discriminability of one or more intrinsic attributes of the speech stimulus. Such an assumption is implicit in any evaluation procedure where speaker recognizability, in the usual sense of the word, is used as a criterion for system performance.

As suggested earlier, however, it is easy to conceive of practical testing situations where the validity of this assumption may be open to question -- particularly where the listener's recognition of a voice can be effected on some basis other than discrimination of the distinguishing acoustical characteristics of the stimulus voice itself.

The indicated solution to this problem is to test directly for the discriminability of intrinsic characteristics which are used by listeners to effect the categorization of voices with respect to some one or other "extrinsic" or arbitrary taxonomy. For, it is perhaps apparent in any case, that our concern with the speaker-recognizability aspect of system performance reduces essentially to a concern for ability to transmit information as to the states or values of the intrinsic attributes or traits which distinguish the voice of one speaker from that of another.

The major purpose of this aspect of the program is, in fact, to determine the nature and number of these "criterial attributes", or traits,

and to formulate practical techniques for evaluating their transmission by an experimental system or device. Accordingly, we shall now consider the problem of developing a voice taxonomy based upon the intrinsic characteristics of voice sounds.

The Problems of Classification

In the case of voices, as in the case of elementary speech sounds, there are several possible bases of classification. Individual differences in terms of size or shape of various structures of the vocal tract offer one basis for a voice taxonomy, though such a taxonomy would have several limitations. Most obvious, perhaps, is the practical difficulty of making the necessary measurements. To the extent, moreover, that individual differences in speech are culturally determined, such a system would not provide an exhaustive characterization of individual differences in speech.

Various physical-acoustic characteristics of individual speech can, of course, provide the basis for the exhaustive classification of voices. But while some characteristics, such as "natural frequency" and "characteristic spectral energy distribution," might be relatively easy to evaluate, other critical characteristics would undoubtedly prove as difficult to measure as to identify. Once the critical physical parameters of individual differences in speech are identified, moreover, there remains the question of their perceptibility for, and use by, a human listener.

Thus, it would appear in the case of voices, as in the case of elementary speech sounds, that a taxonomy or classification system based on the perceived voice characteristics, somewhat analogous to the system of "distinctive features" of Jakobson and Halle (1956), would provide the most generally satisfactory solution to the problem. Among other things, it would provide valuable guidelines for the subsequent development of a voice taxonomy

based on physical-acoustical characteristics. Accordingly, a substantial part of the present program constitutes the initial step toward the development of a perceptual voice taxonomy - based on perceived acoustic traits (PATs). The first question to be treated is that which concerns the most appropriate means of developing a perceptual voice taxonomy.

One time-honored approach to the characterization of individual differences in general is to treat certain highly atypical or pathological conditions as representing endpoints of a continuum along which representatives of the "normal" population may be ordered. A familiar example of this approach is provided by the Minnesota Multiphasic Personality Inventory, which attempts to characterize individual personalities in terms of schizoid, psychasthenic, depressive, and other continua whose endpoints are associated with pathological states. By analogy, perceived "breathiness," "hoarseness," "harshness," etc., suggest themselves as possible continua on which the voices in general may be ordered and, in turn, as potential perceived acoustic traits. Beyond the question, however, of whether these characteristics are sufficiently distributed within the normal population to carry substantial amounts of speaker-identity information, there remains the question of whether they exhaust the possibilities of perceptually-significant, individual differences within the "normal" population of speakers. An affirmative response to this question seems rather unlikely. Whatever the value of these characteristics for purposes of classifying voices, additional dimensions would undoubtedly be required.

The problem, in other words, is to devise an exhaustive catalog of the basic ways in which voices are commonly perceived to differ from each other by the typical listener. Each of these "ways" is, in effect, a potential dimension of a voice taxonomy. Each, in turn, represents a potential

dimension of perceptually useful speaker identity information; this line of reasoning provides the rationale of a practical method of system evaluation from the standpoint of potential speaker recognizability. Specifically, the indicated basis of evaluation is relative capacity for speaker identity information transmittable to listeners via the various dimensions of a voice taxonomy based on perceived acoustic traits.

Three problems then remain. The first concerns the nature and number of perceived acoustic traits which carry potentially significant amounts of speaker identity information. Second is the practical problem of evaluating the speaker identity information received by listeners under a given transmission condition. Third is the practical problem of determining the status of an individual voice with respect to a particular perceived acoustic trait.

A study by Voiers (1964) represents an exploratory treatment of the first of these problems. It serves to demonstrate the usefulness of one method of approach and, at the same time, to provide experimental results of some intrinsic value in relation to the taxonomy problem.

Briefly, the aim of the method used by Voiers is to effect a circumstance in which every conceivable aspect of the typical listener's response to the distinguishing features of individual voices can be isolated and quantified. The "semantic differential method" provides a practical means of achieving this circumstance. Specifically, it involves a situation where listeners use an appropriately designed multi-dimensional rating form to register their perceptions of the distinguishing features of individual voices. An analyses of the invariances of listener response attributable to speaker differences then serves to identify the major dimensions of perceived variability among voices.

The crucial assumption, here, is that a semantic differential rating form does permit an exhaustive characterization of a listener's response to the distinguishing characteristics of individual voices. While such an assumption may prove somewhat difficult to reconcile with intuition, a vast body of experimental data attests to its validity, at least at the pragmatic level.

A typical semantic differential rating form is shown in Fig. 3.1. Using such a form, listeners rate speech samples from each member of a sample of speakers on each of the various semantic continua. Ratings on each semantic continua are averaged for each speaker.

Factor analyses of the inter-correlations among the various "items" or semantic continua serve to reveal the "implicit dimensionality" of the "speaker component" of listener response as represented by the patterns of averages for the total set of semantic continua.

Results from the original study by Voiers indicated that the total variance of ratings on a 49-item form (Fig. 3.1) could be accounted for in terms of only four implicit dimensions or factors. While the need for further research on the number and nature of such factors was clearly apparent, the potential value of the general approach was rather clearly established. This approach, with certain refinements, was thus used in the present program to resolve the issue of the nature and number of elementary dimensions of perceived variability among voices.

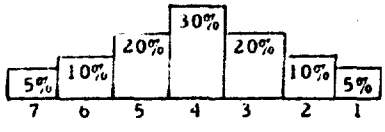
Once the factorial structure of a set of voice ratings data is established, it is possible to replace the set of values representing a speaker's status on the various explicit rating dimensions by a set of values representing his status on the reduced set of implicit or factorial dimensions, each of which is, at least potentially, a perceived acoustic trait. The

Listener _____

Speaker _____

SPEECH RATING FORM IIA

LOUD	7 6 5 4 3 2 1	SOFT
HEAVY	7 6 5 4 3 2 1	LIGHT
BEAUTIFUL	7 6 5 4 3 2 1	UGLY
CLEAR	7 6 5 4 3 2 1	HAZY
FRIENDLY	7 6 5 4 3 2 1	BELLIGERENT
RELAXED	7 6 5 4 3 2 1	TENSE
FAMILIAR	7 6 5 4 3 2 1	STRANGE
COLORFUL	7 6 5 4 3 2 1	COLORLESS
WARM	7 6 5 4 3 2 1	COOL
RISING	7 6 5 4 3 2 1	FALLING
LARGE	7 6 5 4 3 2 1	SMALL
PLEASANT	7 6 5 4 3 2 1	UNPLEASANT
DEFINITE	7 6 5 4 3 2 1	UNCERTAIN
GENTLE	7 6 5 4 3 2 1	VIOLENT
LOOSE	7 6 5 4 3 2 1	TIGHT
WET	7 6 5 4 3 2 1	DRY
RICH	7 6 5 4 3 2 1	THIN
DULL	7 6 5 4 3 2 1	SHARP
MASCULINE	7 6 5 4 3 2 1	FEMININE
RUMBLING	7 6 5 4 3 2 1	WHINING
GOOD	7 6 5 4 3 2 1	BAD
EVEN	7 6 5 4 3 2 1	UNEVEN
CALMING	7 6 5 4 3 2 1	EXCITING
SOFT	7 6 5 4 3 2 1	HARD
ACTIVE	7 6 5 4 3 2 1	PASSIVE
HAPPY	7 6 5 4 3 2 1	SAD
RUGGED	7 6 5 4 3 2 1	DELICATE
FAST	7 6 5 4 3 2 1	SLOW
WIDE	7 6 5 4 3 2 1	NARROW
PLEASING	7 6 5 4 3 2 1	ANNOYING
CONCENTRATED	7 6 5 4 3 2 1	DIFFUSED
REASSURING	7 6 5 4 3 2 1	DISTURBING
SERENE	7 6 5 4 3 2 1	AGITATED
STEADY	7 6 5 4 3 2 1	FLUTTERING
DELIBERATE	7 6 5 4 3 2 1	CARELESS
GLIDING	7 6 5 4 3 2 1	SCRAPING
EASY	7 6 5 4 3 2 1	LABORED
LOW	7 6 5 4 3 2 1	HIGH
SMOOTH	7 6 5 4 3 2 1	ROUGH
OBVIOUS	7 6 5 4 3 2 1	SUBTLE
SIMPLE	7 6 5 4 3 2 1	COMPLEX
MILD	7 6 5 4 3 2 1	INTENSE
NATIVE	7 6 5 4 3 2 1	FOREIGN
FULL	7 6 5 4 3 2 1	EMPTY
POWERFUL	7 6 5 4 3 2 1	WEAK
DEEP	7 6 5 4 3 2 1	SHALLOW
BUSY	7 6 5 4 3 2 1	RESTING
REPEATED	7 6 5 4 3 2 1	VARIED
CLEAN	7 6 5 4 3 2 1	DIRTY



Expected Long-Run Distribution of Ratings

Additional Comments (Other words or phrases which might be used to characterize the sound of this speaker's voice.)

_____

_____

_____

_____

_____

Do Not Write in This Block

Factor I _____

Factor II _____

Factor III _____

Factor IV _____

Factor V _____

Factor VI _____

Factor VII _____

_____

_____

_____

_____

_____

Figure 3.1 A Typical Semantic Differential Rating Form (Osters, 1970)

principle involved here may be somewhat cumbersome when reduced to practice. For various reasons, discussed later, something other than "exact factor scores" may provide the most useful measures of individual status with respect to a given perceived acoustic trait.

It suffices for now, however, to indicate the existence of a means of characterizing individual voices in terms of some limited number of intrinsic attributes or traits. We may turn then to the question of the informational implications of such traits and their uses for purposes of systems evaluation.

The Evaluation of Speaker Identity Information in Perceived Acoustic Traits

Once we have isolated a set of perceived acoustic traits (or other parameters of inter-individual variation in speech) and implemented a means of estimating their values for individual speakers, we may treat the question of their contribution to the speaker recognition process. More specifically, we may turn to the problem of evaluating effects of a particular transmission condition upon the speaker identity information transmitted to a typical listener via a given trait or set of traits. A crucial concept in this connection is the "true value" of a speaker on a particular trait, by which is meant simply the average of the values which would be assigned a speaker by a given population of listeners under a given transmission condition. For, in one sense, the purpose of the method to be described here is nothing more than to evaluate the relative fidelity with which the five coordinates of speakers(in a perceived acoustic trait space)are transmitted by a given system or device. In this connection we may distinguish two types of fidelity: one with respect to the true speaker coordinates for the experimental condition under consideration and the second with respect to true speaker coordinates for the control case. The first of these is thus related to the issue of potential speaker recognizability for a

given system: the possibilities for speaker recognition assuming that listeners are provided the opportunity to familiarize themselves with voices as transmitted or processed by a given system. Conceivably, this potential for recognizability is to some degree independent of the fidelity with which the system preserves the perceptually significant distinguishing features of unprocessed voice sounds.

The second type of fidelity relates to the issue of the recognizability au nouveau of experimentally processed voice samples; i.e., this recognizability is dependent upon familiarity with the characteristics which distinguish them in their normal or unprocessed state.

We cannot, of course, determine the true values of voices on the various perceived acoustic traits, but we can, through analysis of the sampling variation of estimated values, derive various indicants of the relative fidelity with which a given system is capable of transmitting physical correlates of true trait values in any given instance. Several criteria of fidelity would perhaps serve our present purposes equally well but the nature of the subject matter with which we are dealing makes the choice of "information-like" measures particularly appropriate.

Our primary basis of system evaluation from the standpoint of speaker recognizability is the relative amount of speaker identity information contained, on the average, in an estimate (average of listeners ratings) of the true value of a voice on a given PAT. In this connection we accordingly define the quantity:

$$\text{"C"}_{\bar{y}\bar{y}}(m) = 1/2 \log_2 \frac{m\sigma_{\bar{y}}^2 + \sigma_e^2}{\sigma_e^2} = 1/2 \log_2 F(m) = 1/2 \log_2 \frac{1}{1 - r_{\bar{y}\bar{y}}(m)}$$

$\bar{y}$ is the "true value" of a voice on a rating dimension or perceived acoustical trait (PAT) for the control case (unprocessed speech); $\bar{y}$ is a sample or observed value (mean) of a voice on a rating dimension or PAT for the control case.

$\bar{x}$ is the "true value" of a voice on a rating dimension or PAT for an experimental case, e.g., vocoderization; $\bar{x}$ is a sample value of a voice for an experimental case.

F(m) is the ratio, "mean square for speakers"/"mean square for error", for a sample of voices, each rated by n listeners.

$r_{\bar{y}\bar{y}}$ is the estimated product moment coefficient of correlation between the true value of a voice rating and the mean of n listener-ratings of a voice.

(r^2_{xx} (or $r_{\bar{y}\bar{y}}$)) is the coefficient of reliability of a mean of n listener ratings of a voice under a control or experimental condition.

The various expressions given for "C"$_{\bar{y}\bar{y}}$(m) may help to elucidate its information theoretic basis and, in turn, indicate the scope of its applicability.

The first expression given for "C"$_{\bar{y}\bar{y}}$(m) in Eq. 3.1 will perhaps be recognized as analogous to the Shannon measure of channel capacity for the continuous case expressed in terms of bits/sample rather than bits/second. Provision is made, however, for the case in which the sampling unit consists of more than a single observation, i.e., of m listeners' ratings.

The second expression for "C"$_{\bar{y}\bar{y}}$(m) shows its relation to the conventional statistic, F, which provides the means for the practical computation of "C"$_{\bar{y}\bar{y}}$. Where the F-ratio, "mean square for speakers"/"mean square for interaction of speakers and listeners", is employed "C"$_{\bar{y}\bar{y}}$(m) is effectively a measure of "shareable" speaker identity information. Several other ratios are of possible theoretical significance (in relation, for example, to the "non-shareable" speaker identity information received by a single listener)

but have relatively limited implications for the practical problem of systems evaluation.

The rightmost term of Eq. 3.1 provides an expression of "C"$_{\bar{y}\bar{y}}$ which may not be immediately evident from the previous expressions. In this term, r^2 is an intraclass correlation which is equivalent in particular to the average correlation between two independent estimates of $\bar{y}$ made under the same conditions. It can be shown, in turn, that r is an estimate of the coefficient of correlation between <u>estimates</u> of a speaker's value on a given PAT and his <u>true</u> value on the PAT. From this it is perhaps apparent that: Not only may we evaluate the information about $\bar{\bar{y}}$ contained in an estimate, $\bar{y}$, of the same parameter, but we may also extend the principle involved here to evaluate the information about $\bar{\bar{y}}$ contained in estimates of other parameters. In particular, we may evaluate the information about $\bar{\bar{y}}$ (the true value of a speaker for the control case) contained in averaged PAT ratings obtained under an experimental condition.

In this connection we define the quantity

$$\text{"C"}_{\bar{\bar{y}}\bar{x}}(m) = 1/2 \log \frac{1}{1 - r^2_{\bar{\bar{y}}\bar{x}}} = 1/2 \log \frac{1}{1 - \frac{r^2_{\bar{y}\bar{x}}}{r^2_{\bar{y}\bar{y}}}}$$

where: $r_{\bar{\bar{y}}\bar{x}}$ is the coefficient of correlation between a true value, $\bar{\bar{y}}$ and sample value, $\bar{x}$.

$r_{\bar{y}\bar{x}}$ is the observed coefficient of correlation between sample values $\bar{y}$ and $\bar{x}$. (Based on equal observations per sample.)

$r^2_{\bar{y}\bar{y}}$ is the coefficient of reliability of $\bar{y}$.

The uses of "C"$_{\bar{y}\bar{x}}$(m) and various other criteria of system performance are not demonstrated here but will be treated in subsequent reports.

Where the PAT's or other parameters of interest are statistically and experimentally independent, the total amount of speaker identity information contained in estimated values of the various PAT's may be calculated by simple summation of obtained values of "C"$_{\bar{y}\bar{y}}$(m). Where, however, the true values of the various PAT's are correlated in some degree - whether naturally or as a consequence of some experimental treatment - the total information content of a set of PAT's estimates will be something less than the sum of the informational values obtained for the individual PAT's.

On the assumption that errors in the evaluation of the various PAT's are uncorrelated, we may obtain a measure of the total amount of speaker identity information contained in averages of m listener ratings on n PAT's under a given condition by means of the quantity

$$\text{"C"}_{\bar{y}\bar{y}}(m,n) = 1/2 \log_2 |r_{ij}| \prod_{i=1}^{n} F_i(n)$$

where r_{ij} is the coefficient of correlation between corresponding values of $\bar{y}$ (or $\bar{x}$) for the ith and the jth PAT's, and F_i is the ratio, "mean square for speakers"/"mean square for interaction of speakers and listeners," for the ith PAT.

It is perhaps apparent that the values obtained for the various forms of "C" depend not only upon the inherent precision of a PAT estimate, but also upon the size of the sampling unit (i.e., number of listeners) on which such estimates are based. Control of the latter factor is thus essential for purposes of comparative evaluation of speech processing devices. Since the use of a single rating of a voice by a single listener would typically

yield values of less than unity for "C"$=_{\bar{y}\bar{y}}$, the use of a larger, if somewhat arbitrary, "standard sample " is indicated. Accordingly, the standard sampling unit, for all present purposes, is the average of sixteen PAT estimates based on two responses per listener for a crew of eight male listeners.

In view of the relative nature of the various criteria proposed above, the results obtained for any given experimental condition have meaning only in relation to results for a control situation, which differs essentially from the experimental condition only in terms of the experimental parameter of interest.

An essential step in the development of the present method of system evaluation is thus the establishment of "norms" for the various performance criteria. Voice rating experiments with unprocessed speech provide the necessary normative data.

In the following section, two voice-rating studies, involving unprocessed speech, are discussed. Their various results serve to establish a basis for the definition of a set of perceived acoustic traits and also to provide baselines for evaluating the effects of experimental treatment upon the information structure of listener evaluations of voices on the various traits.

The Speaker Identity Information Structure of Multidimensional Voice Ratings

Two major investigations were conducted in an attempt to resolve the issue of the nature and number of perceived acoustic traits and to provide normative data on their speaker identity information content. The investigations are similar in several major respects and, accordingly, provide a basis for the cross validation of several crucial points. However, there are some significant differences in the procedures and materials employed in the two cases, and some of their implications warrant discussion.

In the First Normative Study the stimulus materials were recorded speech samples of 16 male speakers. Each sample consisted of 24 "every day" sentences (Appendix III), spoken at a rate of one sentence per five seconds. These materials were presented to 32 listeners, in groups of eight, by means of a hi-fidelity loud-speaker (Scott S-2). Following a brief exposure to each voice sample, listeners used a 24-item multidimensional rating form to characterize their perceptions of individual voices. The rating form was devised by combining various pairs of previously-used rating scales into single scales. The basis for combination was high correlation between listeners' ratings (Voiers, 1964) on the two scales comprising each pair. One version of the resulting form is presented in Fig. 3.2.

The Second Normative Study was distinguished most significantly, perhaps, by the use of a "relative," rather than an "absolute," rating procedure by the use of a further abbreviated rating form (Fig. 3.3) and by an increase in the size of the speaker sample. These and other features which distinguish the two studies are summarized in Table 3.1.

Generally, the innovations of the second study could be expected, a priori, to enhance the reliability or stability of listeners' ratings and, in turn, to increase the values of the various measures of transmitted speaker-identity information. However, several of these innovations (in particular the increased stimulus-presentation rate and the use of headphones) were motivated primarily by practical considerations which were not necessarily compatible with the dictates of theory.

From the results of the two studies it appears that the combined effects of the various innovations were not altogether favorable. However, further research will be required to isolate these causes. The Second Normative Study was not designed for this purpose, but simply to utilize all

Speaker_____

SPEAKER RATING FORM A

Listener_____

Colorless Monotonous	☐ ☐ ☐ ☐ ☐ ☐ ☐	Colorful Dynamic
Rumbling Low	☐ ☐ ☐ ☐ ☐ ☐ ☐	Whining High
Sharp Shrill	☐ ☐ ☐ ☐ ☐ ☐ ☐	Dull Muffled
Fluttering Unstable	☐ ☐ ☐ ☐ ☐ ☐ ☐	Steady Stable
Thin Empty	☐ ☐ ☐ ☐ ☐ ☐ ☐	Rich Full
Repeated Simple	☐ ☐ ☐ ☐ ☐ ☐ ☐	Varied Complex
Foreign Rare	☐ ☐ ☐ ☐ ☐ ☐ ☐	Native Common
Scraping Rough	☐ ☐ ☐ ☐ ☐ ☐ ☐	Gliding Smooth
Familiar Usual	☐ ☐ ☐ ☐ ☐ ☐ ☐	Strange Unusual
Active Brisk	☐ ☐ ☐ ☐ ☐ ☐ ☐	Passive Dragging
Even Rhythmic	☐ ☐ ☐ ☐ ☐ ☐ ☐	Uneven Unrhythmic
Hazy Uncertain	☐ ☐ ☐ ☐ ☐ ☐ ☐	Clear Definite
Obvious Conspicuous	☐ ☐ ☐ ☐ ☐ ☐ ☐	Subtle Obscure
Shallow Small	☐ ☐ ☐ ☐ ☐ ☐ ☐	Deep Large
Beautiful Clean	☐ ☐ ☐ ☐ ☐ ☐ ☐	Ugly Dirty
Gentle Soft	☐ ☐ ☐ ☐ ☐ ☐ ☐	Violent Hard
Abrupt Jagged	☐ ☐ ☐ ☐ ☐ ☐ ☐	Gradual Rounded
Unpleasant Annoying	☐ ☐ ☐ ☐ ☐ ☐ ☐	Pleasant Pleasing
Fast Busy	☐ ☐ ☐ ☐ ☐ ☐ ☐	Slow Resting
Heavy Masculine	☐ ☐ ☐ ☐ ☐ ☐ ☐	Light Feminine
Friendly Warm	☐ ☐ ☐ ☐ ☐ ☐ ☐	Belligerent Cool
Calm Serene	☐ ☐ ☐ ☐ ☐ ☐ ☐	Excited Agitated
Loud Intense	☐ ☐ ☐ ☐ ☐ ☐ ☐	Soft Mild
Delicate Weak	☐ ☐ ☐ ☐ ☐ ☐ ☐	Rugged Powerful

Figure 3.2 Multidimensional Rating Form Used In The First Normative Study

Speaker_____

SPEAKER RATING FORM III A

Listener_____

Steady Stable	☐☐☐☐☐☐☐☐☐	Fluttering Unstable
Colorless Monotonous	☐☐☐☐☐☐☐☐☐	Colorful Dynamic
Foreign Rare	☐☐☐☐☐☐☐☐☐	Native Common
Rumbling Low	☐☐☐☐☐☐☐☐☐	Whining High
Unpleasant Annoying	☐☐☐☐☐☐☐☐☐	Pleasant Pleasing
Gradual Rounded	☐☐☐☐☐☐☐☐☐	Abrupt Jagged
Loud Intense	☐☐☐☐☐☐☐☐☐	Soft Mild
Passive Dragging	☐☐☐☐☐☐☐☐☐	Active Brisk
Excited Agitated	☐☐☐☐☐☐☐☐☐	Calm Serene
Gliding Smooth	☐☐☐☐☐☐☐☐☐	Scraping Rough
Fast Busy	☐☐☐☐☐☐☐☐☐	Slow Resting
Beautiful Clean	☐☐☐☐☐☐☐☐☐	Ugly Dirty
Feminine Light	☐☐☐☐☐☐☐☐☐	Masculine Heavy
Familiar Usual	☐☐☐☐☐☐☐☐☐	Strange Unusual
Clear Definite	☐☐☐☐☐☐☐☐☐	Hazy Uncertain
Uneven Irregular	☐☐☐☐☐☐☐☐☐	Even Regular

Figure 3.3 Multidimensional Rating Form Used In The Second Normative Study

Table 3.1 Materials and Methods of the Normative Voice-Rating Studies

	First Normative Study	Second Normative Study
Size of Speaker Sample	16 adult males	24 adult males (including 16 from Normative Study No. 1)
Stimulus Materials	24 "every day" sentences	16 "every day" sentences
Stimulus Presentation Rate	1 sentence/five seconds	1 sentence/four seconds
Manner of Stimulus Presentation	Scott S-2 loudspeaker	PDR-8 headphones (diotic)
Frequency-passband	60-7500 H_z	200-4000 H_z
Number of Rating Dimensions	24	16 (selected from original twenty-four from Normative Study No. 1)
Number of Rating Categories	7	9
Identification of Individual Speakers	Identifying phrase spoken by each speaker (e.g., "This is Speaker No. 7")	Identifying phrase spoken by announcer (e.g., "Now you will hear the voice of Speaker No. 7")
Rating Standard	"Subjective impression of the typical voice," formed on the basis of preliminary exposure to all voices	The voice of the announcer (a "neutral voice" selected on the basis of results of Normative Study No. 1)
Listeners	Sixteen male and sixteen female college students (Brandeis University)	Thirty-two male college students (Tufts University)
Relevant Listener Experience	None	Average of approximately 12 hours as subjects for intelligibility tests

available theory and experimental results to the end of expediting development of the optimal procedure. That this attempt was not completely successful does not seriously detract from the value of the obtained results for purposes of comparative evaluation of the effects of various experimental treatments of the speech signal. Let us now consider some of the more significant implications of the studies considered jointly and individually.

Both studies provide results which bear upon the fundamental issue of the dimensionality of voice ratings or upon the question of the nature and number of independent perceived acoustic traits (PAT's). Results were analyzed by the same methods in both cases. Mean ratings received by speakers on various rating dimensions were subjected to factor analysis by the method of principle components. Coefficients of reliability for mean ratings (intra-class correlations) were used as initial estimates of item communalities to ensure exhaustive factorial representation of the variance attributable to speakers. The initial factorial axes were rotated to satisfy a varimax criterion of simple structure.

Final factor loadings for the case of the First Normative Study are presented in Table 3.2. Several aspects of these results are worthy of comment.

First is the number of factors revealed by the analysis. As in the case of the earlier study by Voiers (1964), four factors suffice to account for essentially all of the rating variance attributable to generally-perceived differences among voices. Moreover, three of the present factors correspond quite closely to factors previously identified by Voiers. Factor I is thus labeled pitch-magnitude.

The second factor is numbered II-III in view of the fact that it is defined by items which have served in other studies to define two factors.

Table 3.2 Final Factor Loadings for Absolute Ratings of the Unprocessed Speech: First Normative Study

Item	Factor I	Factor II-III	Factor IV	Factor V	Communality	Reliability
Rumbling-Whining High-Low (2)*	.87	-.391	.247	-.067	.98	.98
Masculine-Feminine Heavy-Light (20)	.97	.099	.198	.014	.99	.98
Loud-Soft Intense-Mild (23)	.26	.933	-.195	-.012	.98	.97
Scraping-Gliding Rough-Smooth (8)	.03	.795	-.541	.111	.94	.95
Active-Passive Brisk-Dragging (10)	.09	.960	-.236	-.033	.98	.98
Fast-Slow Busy-Resting (19)	-.12	.902	-.372	-.112	.98	.99
Beautiful-Ugly Clean-Dirty (15)	.34	-.282	.835	.071	.91	.88
Clear-Hazy Definite-Uncertain (12)	.31	.868	.291	.044	.94	.89
Foreign-Native Rare-Common (7)	-.28	-.113	-.515	.646	.78	.78
Strange-Familiar Unusual-Usual (9)	-.48	-.155	-.703	.390	.89	.89
Steady-Fluttering Stable-Unstable (4)	.51	-.208	.818	.061	.98	.96
Colorful-Colorless Dynamic-Monotonous (1)	.36	787	.089	.372	.91	.93
Pleasant-Unpleasant Pleasing-Annoying (18)	.63	.275	.696	026	96	96
Abrupt-Gradual Jagged-Rounded (17)	-.15	832	- 513	.121	.99	.97

* Item No. on Original Rating Form

Table 3.2 Final Factor Loadings for Absolute Ratings of the Unprocessed Speech: First Normative Study (cont.)

Item	Factor I	Factor II-III	Factor IV	Factor V	Communality	Reliability
Excited-Calm Agitated-Serene (22)	-.19	.873	-.428	-.059	.99	.99
Even-Uneven Regular-Irregular (11)	.28	-.681	.647	-.095	.98	.93
Sharp-Dull Shrill-Muffled (3)	-.52	.815	-.170	-.009	.97	.97
Thin-Rich Empty-Full (5)	-.92	-.008	-.371	-.106	.99	.91
Repeated-Varied Simple-Complex (6)	-.40	-.640	.204	-.556	.92	.89
Obvious-Subtle Conspicuous-Obscure (13)	-.08	.875	.155	.092	.80	.84
Shallow-Deep Small-Large (14)	-.94	-.014	.124	-.073	.98	.98
Gentle-Violent Soft-Hard (16)	.23	.950	-.216	.022	.99	.97
Friendly-Belligerent Warm-Cool (23)	.05	.570	.70	-.200	.96	.93
Rugged-Delicate Powerful-Weak (24)	.93	.274	-.02	-.021	.98	.98

Specifically, the loudness-roughness and animation-rate dimensions of the other studies appear to be superimposed in this case. Both those items having a prima facie relation to speech rate and those items normally related to the intensive aspects of speech are heavily loaded on this factor. However, many of these items have substantial loadings on Factor IV, which further complicates the problem of interpretation.

Factor IV appears to represent a beauty dimension which has emerged in several other instances. Here, as in other instances, it is rather poorly defined, however.

Factor V represents a new development in relation to the issue of the number of elementary perceived acoustic traits. The two items having highest loadings on this factor would seem to have in common a connotation of normality or perhaps naturalness. Accordingly, this factor is labeled normality.

All in all, these results present a somewhat ambiguous picture concerning the nature and number of elementary perceived acoustic traits. It seemed possible that some clarification of this picture could be accomplished through arbitrary rotation of the factorial axes (the varimax method of rotation is by no means infallible nor optimally suited to all purposes). However, a cursory exploration of this possibility did not prove particularly fruitful. A solution to the dilemma posed by these results is not immediately apparent, though the results of the second normative study throws some light on the matter.

Table 3.3 presents the results of the factor analysis of voice rating data from the second normative study. Several aspects of these results are of interest, particularly in view of the procedural innovations which distinguish the study from the first normative study. Especially noteworthy is the presence of a fifth factorial dimension. Rarely, if ever, in the literature of the "semantic" differential method are examples found where more

Table 3.3 Final Factor Loadings for Relative Ratings of Unprocessed Speech: Second Normative Study

Item	Factor I Pitch-Magnitude	Factor II Loudness-Roughness	Factor III Animation-Rate	Factor IV Clarity-Beauty	Factor V Normality	Communality	Reliability
Rumbling-Whining High-Low (4)*	.92	.04	.37	.03	-.02	.96	.99
Masculine-Feminine Heavy-Light (13)	.94	-.16	-.26	.13	-.08	.99	.98
Loud-Soft Intense-Mild (7)	.20	.90	.19	.22	.10	.95	.94
Scraping-Gliding Rough-Smooth (14)	.16	.79	.34	-.24	.29	.91	.86
Active-Passive Brisk-Dragging (8)	-.38	.54	.69	.17	.10	.96	.96
Fast-Slow Busy-Resting (11)	-.52	.54	.63	-.04	.02	.96	.97
Beautiful-Ugly Clean-Dirty (12)	-.17	-.31	-.02	.78	-.31	.83	.79
Clear-Hazy Definite-Uncertain (15)	.04	.49	-.13	.75	-.13	.84	.83
Foreign-Native Rare-Common (3)	.03	.12	-.04	-.02	.91	.85	.85
Strange-Familiar Unusual-Usual (14)	-.14	-.08	.19	-.54	.74	.91	.87
Steady-Fluttering Stable-Unstable (1)	.62	-.08	-.61	-.03	-.41	.93	.90
Colorful-Colorless Dynamic-Monotonous (2)	.15	.16	.59	.66	.30	.92	.92
Pleasant-Unpleasant Pleasing-Annoying (5)	.56	-.19	-.08	+.68	-.33	.93	.89
Abrupt-Gradual Jagged-Rounded (6)	-.34	.75	.46	-.08	.27	.97	.93
Excited-Calm Agitated-Serene (9)	-.53	.55	.63	-.05	.04	.98	.97
Even-Uneven Regular-Irregular (16)	.36	-.23	-.19	.15	-.31	.87	.92

*Item No. in Original Rating Form

than four dimensions (three is most common) are required to account for the "stimulus" or "concept" component of variance in listeners' responses. This suggests that the introduction of a reference voice has the desired effect of enhancing the sensitivity of the method to qualitative as well as quantitative differences among voices.

As in previous studies with the voice rating method, Factor I is designated pitch magnitude on the basis of the high loadings exhibited by items containing such terms as "low", "rumbling", "masculine", etc. It is perhaps the most univocally defined of all factors in the sense that the items most highly loaded on this factor are at once negligibly loaded on all others, although several items, which are clearly identified with other factors, exhibit substantial loadings on this factor. As suggested in an earlier report (Voiers, 1964) it seems quite likely that ratings on this dimension are correlated in some degree with the speaker's natural or average pitch frequency. Some results reported by Holmgren (1964) are thus of interest in this connection. For a sample of 10 voices, that investigator obtained averaged pitch frequencies (as measured by a vocoder pitch extractor) and ratings on items similar to those which define Factor I. Coefficients of correlation between pitch frequency and mean ratings on the various items were of the order of 0.80. In view of the small size of Holmgren's speaker sample, these results merit only qualified acceptance pending further research on the issue. However, the results of a somewhat different approach may have some interest in relation to this issue, and tend to support Holmgren's findings.

It has been noted that a speaker's "natural frequency" tends to be systematically related to the lowest tone that he can vocalize. In view of the ease with which the latter variable can be evaluated, it seemed worthwhile to examine its relation to voice ratings. For this purpose, average ratings

received by each speaker on items 4 and 13 of Rating Form IIIA were averaged to obtain an estimate of each speaker's status with respect to the pitch-magnitude dimension. The lowest tone which each speaker could sing was also determined. In Figure 3.4 the frequency of each speaker's lowest tone is plotted against his averaged rating. From this scattergram it appears that the rating values do not depend in a simple, linear manner upon lowest-tone frequency. For, while low values of the latter tend to be rather consistently associated with ratings of "rumbling", "masculine," etc., high lowest-tone frequencies do not necessarily ensure ratings toward the "whining-feminine" end of the scale. Speakers with relatively high lowest-tones may be perceived to have low, masculine voices — evidently on the basis of acoustical characteristics other than natural frequency. While based on a different class of stimulus materials, some results from Solomon's study of passive sonar sounds (1955) are consistent with this hypothesis. Ratings on a "magnitude" dimension were found to be most highly correlated with energy variations in the 300-600 Hz range, though some correlation existed with variations in the lower regions of the frequency scale. Clearly, more research is in order at this point.

Factor II appears to be the Loudness-Roughness factor of the earlier study and is fairly well defined by items nominally associated with these characteristics. Substantial loadings for certain other items suggested, however, that the stimulus correlates of judged loudness will not be found only in the intensive aspects of the speech signal. In the 1964 study of Voiers an attempt was made to preserve individual differences in natural speech levels in stimulus materials presented to the listeners. In more recent studies, the attempt was made to present all voices at approximately the same level. A comparison of results for items nominally pertaining to intensive differences in voices shows, if anything, that control of individual differences in speech

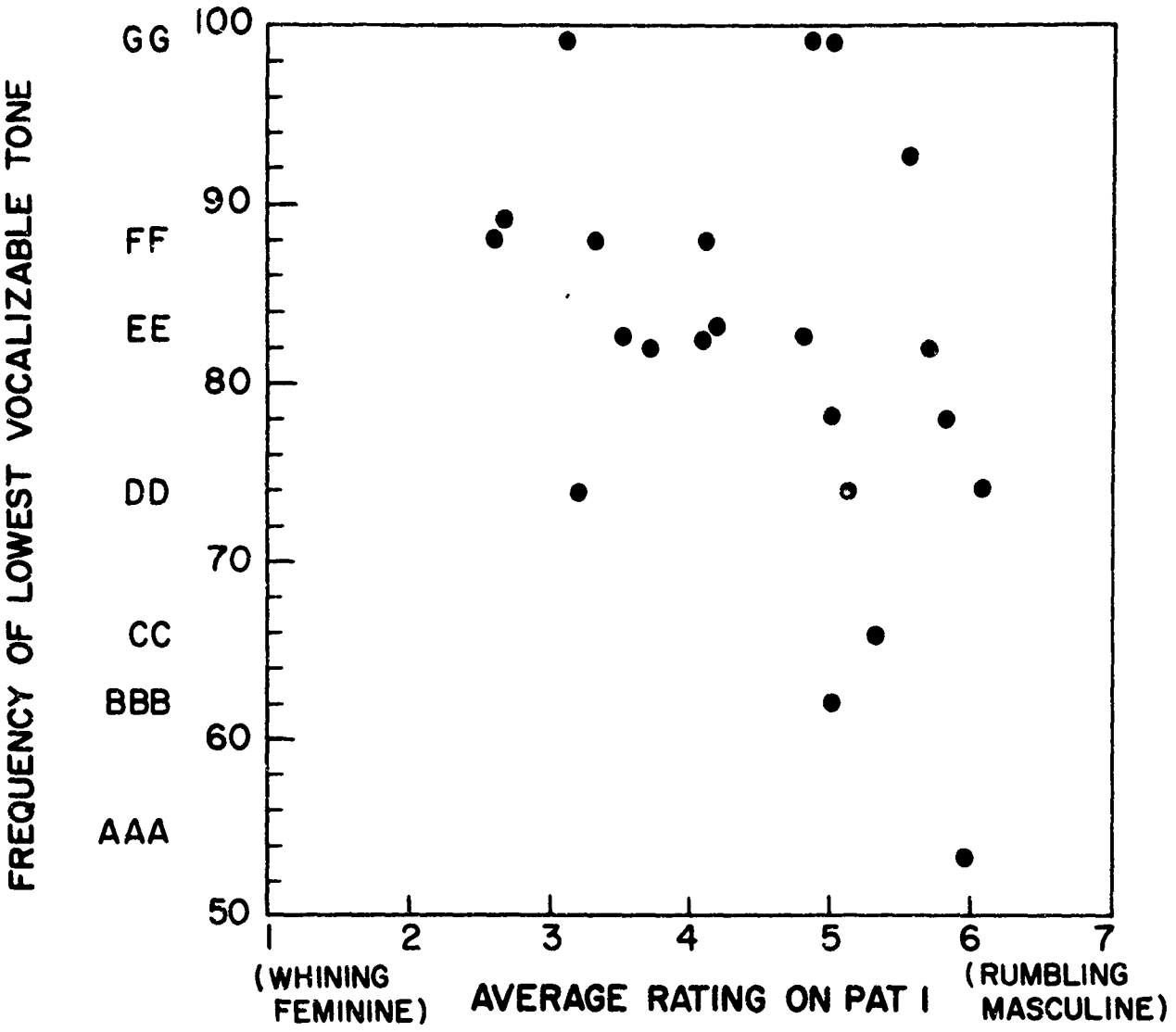


FIG. 3.4 Scattergram of lowest vocalizable tone vs average rating on Pat I (pitch-magnitude).

levels actually enhances the reliability and, hence, the speaker-identity information content of voice ratings on those items having intensive connotations. The results of Solomon's study (1959) suggest that perceived intensity, or loudness, of complex sounds is a function of the concentration of energy in the range from 15 - 2400 Hz relative to the range above 2400 Hz. These results are roughly consistent with the hypothesis that the stimulus correlate of inherent loudness resides in the voiced portions of speech or perhaps in the ratio of voiced-to-unvoiced energy. Holmgren (1964) reports high correlation between intensive judgements and measures both of voiced and unvoiced energy. Somewhat higher values were observed in the case of the former, however. Again, it is therefore appropriate that we reserve judgement as to the physical correlates of a perceived acoustic trait pending the completion of further relevant research.

The label animation-rate appears most appropriate for the case of Factor III, which has a well-defined counterpart in the results of all voice rating studies conducted thus far. Here, as in earlier studies, items loading heavily on this factor tend also to have substantial loadings on certain other factors - in particular, Factor II. In itself, this would seem to suggest that the stimulus correlate of perceived rate-of-speaking is something more obscure than simply the actual rate at which the speaker enunciates speech sounds. This issue is of particular significance when considered in light of the sorts of transformations of the voice signal typically performed by modern speech compression devices. To the extent that speech rate per se is the stimulus correlate of ratings on this dimension, one would predict vocoderization to have little effect upon the amount of speaker identity information received by listeners via this perceptual dimension.

To help resolve this issue, therefore, the correlation between speaker "scores" on the two "rate" items and a simple measure of speech-rate

(total time required to enunciate the 16 sentences used in the second normative study) was computed. A value of only .58 was obtained. Moreover, adjustment for the unreliability of the rating data increased this value by only 0.02. The implication of these results, therefore, is that a substantial amount of systematic variation in this dimension is unaccounted for by speech rate per se. Some results by Holmgren (1964) are consistent with this proposition. That investigator found voice ratings on the "fast-slow" and "busy-resting" continua to be virtually uncorrelated with the speaker's actual rate of enunciation, though they were correlated with a measure of "average" amplitude of voiced sounds. The latter finding is of some interest in light of the indication, described earlier, to the effect that Factors II and III are not consistently orthogonal. Several studies of vocoded speech, to be described at another point, lend further support to the hypothesis of complexly-determined judgements on the animation-rate dimension. However, further research will be required to resolve this issue.

Factor IV corresponds most closely to the dimension which Voiers has previously labeled "clarity". As in the previous study, however, a number of the items with high loadings on this factor have a fairly pronounced "aesthetic-evaluative" connotation. To the extent that this is the case, one would reasonably expect a substantial amount of variability among listeners in their responses to individual voices. Generally, one would, in turn, expect the speaker identity information in this dimension of listener response to be relatively small, which is consistently confirmed by results described later in this report.

Factor V finds some precedence in the results of Solomon's work (1959) if only in the sense that a factor defined by similar rating-scales (e.g., "familiar-strange") emerged from the analysis of listener responses

to sonar sounds. While Solomon was able to delineate to some degree the stimulus correlates of this dimension of perceived variability, any attempt to generalize his results to the case of voices would hardly seem warranted. A priori, there is no basis for expecting that "strange-sounding sonar signals" would have anything in common, perceptually or physically, with "strange-sounding voices."

Consider now some of the broader implications of the various results described thus far. Of most immediate interest are implications for the dimensionality issue. With regard to this issue the results present a picture which, though ambiguous in some respects, contains some significant invariances. Thus, while it is clear that the dimensionality of listeners' perceptual reactions to voices is, to some extent, dependent on the experimental method employed, the number and nature of the major dimensions of listener response are rather clearly indicated. They appear, moreover, to be relatively independent of the method and materials employed. Thus a pitch-magnitude or similar factor has also emerged in studies of other types of acoustical stimulus materials and even in studies involving non-acoustical stimulus materials (e.g., Elliot and Tannenbaum, 1963). Factors similar to the loudness-roughness and animation-rate have also appeared consistently, though they do not always appear to represent orthogonal dimensions of listener-response to voices.

While not so univocally defined, a factor similar to the clarity-beauty factor of the second normative study has been observed consistently. In at least three respects, therefore, the essential structure of listener response to acoustical and other stimuli appears to be quite stable. It is of some interest that Factors I, II, and IV appear to correspond rather closely to the potency, activity, and evaluative dimensions which have emerged in semantic

differential studies in a diversity of situations. Factor V finds less precedence in the results of research with the semantic differential, though under some circumstances there would be no reason for expecting it to appear.

On the basis of such insights concerning the factorial structure of voice ratings, we may now consider the issue of defining and measuring perceived acoustic traits. In principle, each of the factorial dimensions revealed in one or more of the experiments discussed above is potentially a perceived acoustic trait, and, as such, may conceivably correspond to a crucial dimension of system performance. However, several things complicate the task of reducing principle to practice here.

In addition to the degree of unpredictability which exists regarding the emergence of certain factors, there is also the question of their orthogonality. While under some circumstances the various factors appear to represent statistically independent aspects of listener response to voices, under others they appear to be correlated in greater or lesser degree. Statistical independence does not guarantee experimental independence in any case. In several instances, moreover, the configuration of item loadings with respect to orthogonal dimensions of listener response may be such that no item, or groups of items, is uniquely associated with each factor. Thus a reasonably "pure" measure of a listener's response to a given factorial dimension cannot be obtained without recourse to relatively complicated computational procedures utilizing response data from a large number of items. Ideally, one would hope to obtain factorially pure measurements of a voice from listener-response data for a single item or, at most, from a simple average of data for a small number of items. At our present level of understanding this ideal cannot be realized if we retain the orthogonality requirement for perceived acoustic traits used as the basis for the classification of voices. Noting, however,

that the degree of orthogonality of the major factorial dimensions appear in any case to be somewhat variable, we may raise the question as to whether it is feasible, or even desirable, to retain the requirement of orthogonality among perceived acoustic traits. Several considerations, both theoretical and practical, suggest that it is not.

Theoretically, there appears the possibility that changes in the degree of orthogonality among perceived voice characteristics may, themselves, constitute potential criteria of system performance. Thus, an orthogonal, non-redundant taxonomy may, in some cases, be insensitive to certain significant changes in the structure of the perceptually significant speaker identity information in transmitted speech.

On the practical side, the relaxation of the orthogonality requirement greatly simplifies the problems of defining and measuring perceived acoustic traits. Somewhat arbitrarily, therefore, we have chosen to identify each of five PAT's with two items from the voice rating form such that the average of the ratings received by a speaker on the appropriate pair of items will represent his "value" on a given PAT. The selection of items to be used in evaluating a given PAT is based on three criteria.

One criterion is the pattern of factor loadings which is typical of each of the items over the course of experiments conducted thus far. Other things being equal, those items are chosen which exhibit the highest loadings on some one of the five factors revealed by the second normative study and in which a similar factor has emerged.

A second criterion relates to the manner in which the selected items are distributed in the five-factor space. As far as possible, they are selected to "bracket" this space in all dimensions.

A third criterion is based on the reliability and, hence, the informational characteristics of an individual item considered in isolation. Satisfaction of the first two criteria presupposes satisfaction of the third to a fairly high degree in that high communalities and/or loadings on individual items cannot occur without high reliability. All other things equal, however, two items may differ in reliability. The items having the higher reliability are preferred for present purposes.

The coefficients of reliability of speaker means for the various items are shown for the first and second normative studies in the last columns of Tables 3.2 and 3.3, respectively.[1] The F-ratios from which they were derived are shown in Tables 3.4 and 3.5. For present purposes, the "standard crew" is composed of eight male listeners, each of whom rates each voice twice on each rating dimension. Averaged in the appropriate manner these ratings yield a given speaker's "score" on a particular PAT. It is then the average amount of speaker identity information in such scores, "C"$\bar{=}_{\bar{y}\bar{y}}(8,1)$, that serves as our basic figure of merit for evaluating a given transmission condition from the standpoint of potential speaker recognizability.

Values of "C"$\bar{=}_{\bar{y}\bar{y}}(8,1)$ for each PAT were averaged for the four eight-member crews used in each of the normative studies, and the obtained values used as the basis for comparative evaluation of various experimental treatments or transmission conditions.

Depending upon the procedures and materials used in a given experimental evaluation, results from the First or Second Normative Study are used as the bases for comparative evaluation.

---

[1] These coefficients, intraclass correlations, are derived from the formula $(F-1)/F$, where F is the ratio formed by the mean square for speakers and by the mean square for interaction of listeners and speakers.

All items selected were used in connection with both normative studies, though it will be recalled that their various relations to the factor structure revealed by the First Normative Study are necessarily somewhat different than in the case of the Second Normative Study.

Specifically, each of five perceived acoustic traits is identified with a selected pair of items, as follows:

PAT I (Pitch-Magnitude)

Rumbling - Whining
Low - High

Masculine - Feminine
Heavy - Light

PAT II (Loudness - Roughness)

Loud - Soft
Intense - Mild

Scraping - Gliding
Rough - Smooth

PAT III (Animation Rate)

Active - Passive
Brisk - Dragging

Fast - Slow
Busy - Resting

PAT IV (Clarity-Beauty)

Beautiful - Ugly
Clean - Dirty

Clear - Hazy
Definite - Uncertain

PAT V (Normality)

Foreign - Native
Rare - Common

Strange - Familiar
Unusual - Usual

Table 3.4 Summary of Analysis of Variance Results for Absolute Ratings of Unprocessed Voice Samples: First Normative Study
F-Ratios for Speaker Effect

Item	Group 1A	Group 1B	Group 1C	Group 1D	Averaged Groups	Pooled Groups
Rumbling-Whining High-Low (2)*	11.54	7.10	14.03	15.64	12.08	42.17
Masculine-Feminine Heavy-Light (20)	17.40	13.76	19.15	19.84	17.53	58.11
Loud-Soft Intense-Mild (23)	9.80	11.14	10.16	14.82	11.48	39.71
Scraping-Gliding Rough-Smooth (8)	4.23	6.88	4.60	5.65	5.34	19.89
Active-Passive Brisk-Dragging (10)	9.51	15.05	15.24	15.38	13.79	55.54
Fast-Slow Busy-Resting (19)	21.00	16.28	17.37	14.63	17.07	65.29
Beautiful-Ugly Clean-Dirty (15)	4.28	2.93	1.87	3.27	3.09	8.64
Clear-Hazy Definite-Uncertain (12)	1.71	2.91	2.31	4.96	2.97	9.37
Foreign-Native Rare-Common (7)	.89	2.09	2.98	1.52	6.37	4.65
Strange-Familiar Unusual-Usual (9)	2.81	3.03	3.27	3.23	3.09	9.17
Steady-Fluttering Stable-Unstable (5)	6.20	6.77	7.77	5.83	6.64	25.83
Colorful-Colorless Dynamic-Monotonous (1)	4.41	6.88	3.53	4.14	4.74	15.02
Pleasant-Unpleasant Pleasing-Annoying (18)	8.89	8.70	6.77	4.96	7.33	26.13
Abrupt-Gradual Jagged-Rounded (17)	9.41	10.11	7.91	13.32	10.19	38.93
Excited-Calm Agitated-Serene (22)	12.69	16.10	16.67	24.93	17.57	69.46
Even-Uneven Irregular-Regular (11)	4.19	4.62	3.26	4.48	4.11	14.78

*Item No. on Original Rating Form

Table 3.1 Summary of Analysis of Variance Results for Absolute Ratings of Unprocessed Voice Samples: First Normative Study(Cont.)

Item	Ratios for Speaker Effect				Averaged Groups	Pooled Groups
	Group 1A	Group 1B	Group 1C	Group 1D		
Sharp-Dull Shrill-Muffled (3)	10.44	8.65	7.87	9.65	9.15	31.23
Thin-Rich Empty-Full (5)	10.96	8.41	11.34	9.85	10.14	37.85
Repeated-Varied Simple-Complex (6)	4.35	3.21	2.68	3.22	3.37	9.17
Obvious-Subtle Conspicuous-Obscure (13)	4.30	1.92	1.75	3.96	2.98	4.98
Shallow-Deep Small-Large (14)	12.44	9.18	18.91	13.84	13.37	46.06
Gentle-Violent Soft-Hard (16)	7.74	13.05	8.51	9.86	9.79	37.69
Friendly-Belligerent Warm-Cool (23)	5.43	5.99	2.73	3.66	4.45	13.79
Rugged-Delicate Powerful-Weak (24)	20.58	13.61	11.10	11.84	14.28	48.38

Table 3.5 Summary of Analysis of Variance Results for Relative Ratings of Unprocessed Voice Samples: Second Normative Study

F-Ratios for Speaker Effect

Item	Group 2A	Group 2B	Group 2C	Group 2D	Averaged Group	Pooled Group
Rumbling-Whining Low-High (4)*	14.46	12.19	12.03	15.83	13.88	48.69
Masculine-Feminine Heavy-Light (13)	14.97	10.67	14.45	18.26	14.59	53.19
Loud-Soft Intense-Mild (7)	3.98	5.68	5.90	4.12	4.67	14.45
Scraping-Gliding Rough-Smooth (10)	2.10	3.20	2.02	3.15	2.62	7.77
Active-Passive Brisk-Dragging (8)	3.79	7.38	7.86	9.75	6.95	23.11
Fast-Slow Busy-Resting (11)	7.49	11.13	8.74	9.43	9.20	30.33
Beautiful-Ugly Clean-Dirty (12)	1.48	2.21	2.02	2.29	2.00	7.27
Clear-Hazy Definite-Uncertain (15)	2.49	2.25	2.55	2.96	2.31	5.53
Foreign-Native Rare-Common (3)	1.63	2.07	3.59	2.02	2.51	6.68
Strange-Familiar Unusual-Usual (14)	1.48	2.42	3.36	4.59	2.96	7.27
Steady-Fluttering Stable-Unstable (1)	3.79	2.31	2.29	4.59	3.29	9.84
Colorful-Colorless Dynamic-Monotonous (2)	3.95	3.87	4.18	3.00	3.53	12.46
Pleasant-Unpleasant Pleasing-Annoying (5)	2.63	3.45	2.23	2.83	2.79	9.12
Abrupt-Gradual Jagged-Rounded (6)	3.94	4.24	3.93	4.30	4.11	13.98
Excited-Calm Agitated-Serene (9)	6.64	8.33	8.18	10.65	8.49	23.23
Even-Uneven Irregular-Regular (16)	2.54	2.65	1.68	3.07	2.00	7.52

*Item No. on Original Rating Form

In all cases discussed below, it is on the basis of these five PAT's that comparative evaluation of the speaker identity information structure of transmitted speech is estimated. In this connection, it is of interest, first, to know the effects of averaging data for the various item-pairs. Given that the systematic, i.e., "speaker," components of the variance of rating on items of a pair are highly correlated, we should expect the F-ratio and, in turn, information content of the averaged items to be somewhat higher than for the case of individual items. Tables 3.6 and 3.7 present these ratios for the first and second normative studies, respectively. A comparison of these with F-ratios for the individual items involved (Tables 3.4 and 3.5) reveals that our expectations are confirmed for PAT's I, II, and III, though the results for PAT's IV and V are not so impressive. The implication in this latter case is that the speaker components of the items involved are not highly correlated, at least in relation to the correlations between the corresponding error components.

Tables 3.8 and 3.9 show for each PAT the average amount of speaker identity information for speaker means [$"C"\overline{\overline{}}_{yy}(m,n)$] based on ratings by various numbers of listeners for the first and second normative studies, respectively. Also shown for each case is $"C"\overline{\overline{}}_{yy}(m,5)$, the total amount of speaker identity information transmitted via the five PAT's for various groups of listeners.

The similarities between the first and second normative studies in terms of both the structure (i.e., distribution of information over PAT's) and total amount of speaker identity is particularly striking in view of the procedural differences between the two studies and the differences in factor analytic results.

Among other things, these results provide us with a set of standards for use in evaluating the effects of various experimental treatment upon the

Table 3.6 Results of Analysis of Variance for Selected Item Combinations: First Normative Study
F-ratio for Speaker Effect

Group	PAT I(2,20)* Pitch-Mag.	PAT II(8,23) Loud-Rough	PAT III(10,19) Anim.-Rate	PAT IV(12,15) Clar.-Beauty	PAT V(7,9) Normality
1A	17.17	9.38	17.46	1.84	1.79
1B	10.71	13.11	19.40	3.18	3.09
1C	22.47	13.03	22.03	2.29	3.50
1D	25.09	11.55	13.32	3.23	2.68
Pooled	58.22	12.51	77.46	7.81	9.16

*Numbers in parenthesis identify items from Speaker Rating Form A which were used to estimate PAT values for each speaker.

Table 3.7 Results of Analysis of Variance for Selected Item Combinations: Second Normative Study
F-ratio for Speaker Effect

Group	PAT I(4,13)* Pitch-Mag.	PAT II(7,10) Loud-Rough	PAT III(8,11) Anim.-Rate	PAT IV(12,15) Clar.-Beauty	PAT V(3,14) Normality
2A	13.7	3.19	6.63	2.24	1.63
2B	14.30	5.27	12.21	2.93	2.70
2C	16.57	4.19	9.51	2.73	1.18
2D	20.43	5.14	11.31	3.00	5.02
Pooled	65.73	14.17	33.84	6.41	8.66

*Numbers in parenthesis identify items from Speaker Rating Form IIIA which were used to estimate PAT values for each speaker.

Table 3.8 Structure of Speaker Identity Information: Absolute Voice Ratings of Unprocessed Speech: First Normative Study

Group	"C"$_{\bar{x}}$(...,n) ... Bits: PAT I (2,20) Pitch-Mag.	PAT II (5,23) Loud-Rough	PAT III (10,15) Anim.-Rate	PAT IV (12,17) Clar.-Beauty	PAT V (7,9) Normality	SUM	"C"$_{\bar{x}}$(m,n)	R^2
1A	2.07	1.60	2.07	.11	.47	6.62	1.20	2.42
1B	1.65	1.83	2.13	.82	.81	7.21	3.98	2.26
1C	2.23	1.82	2.21	.62	.90	7.78	5.37	2.41
1D	2.32	1.90	2.09	.82	.70	7.83	5.13	2.60
Grps. Avg'd	2.07	1.78	2.12	.67	.71	7.35	1.02	2.73
Grps. Pooled ("C"$_{\bar{x}}$(32,n))	2.99	2.70	3.13	1.47	1.51	1.81	6.18	3.33

[1] Numbers in parentheses identify items from Speaker Identity Form A which were used to estimate PAT values for each speaker.

[2] Redundancy: "Sum" - "C"$_{\bar{x}}$(m,n)

Table 3.9 Structure of Speaker Identity Information in Relative Voice Ratings of Unprocessed Speech: Second Normative Study

Group	"C"$_{\bar{x}}$ in Bits: PAT I (1,13) Pitch-Mag.	PAT II (7,10) Loud-Rough	PAT III (8,11) Anim.-Rate	PAT IV (12,15) Clar.-Beauty	PAT V (3,11) Normality	SUM	"C"$_{\bar{x}}$(m,n)	R^2
2A	2.10	.82	1.35	.37	.35	5.19	3.82	1.37
2B	1.91	1.18	1.78	.75	.71	6.33	4.40	1.93
2C	2.01	1.02	1.61	.73	1.02	6.39	4.86	1.53
2D	2.16	1.16	1.73	.78	1.11	6.97	5.30	1.67
Grps. Avg'd	2.05	1.04	1.62	.71	.80	6.23	4.72	1.70
Grps. Pooled ("C"$_{\bar{x}}$(32,n))	3.01	1.90	2.51	1.33	1.55	10.33	8.73	1.33

[1] Numbers in parenthesis identify items from Speaker Rating Form IIIA used to estimate PAT values for each speaker.

[2] Redundancy

amount and structure of speaker identity information transmitted to listeners via the five PAT's.

The average amount of speaker identity information contained in a speaker's "score" on a given PAT will, of course, depend upon the number of listener's ratings involved. For present purposes a "standard crew" of eight listeners is employed. Some typical experimental results may serve to demonstrate the sensitivity and validity of the voice rating method. They also serve to provide some useful insights concerning the implications of selected vocoder techniques for potential speaker recognizability.

Table 3.10 summarizes the results of several experiments in which groups of eight listeners used Speaker Rating Form A to rate the voices of 16 speakers on 24 semantic continua. But, for the manner in which the stimulus materials were processed before presentation to the listeners, the procedures were identical to those of the First Normative Study. Estimates of the speaker-identity information transmitted via five PAT's are shown for six cases of vocoder speech, along with comparative results from the first normative study. In the first three cases, the stimulus materials were presented to listeners by means of a loudspeaker, as in the case of the First Normative Study. In the last three, PDR-8 headphones were used. The same master tapes were used in preparing all of the recorded material. Before attempting to interpret these results, we should consider the question of statistical significance.

While a practicable method for testing the significance of differences between values of "C"$\overline{xx}$ is yet to be derived, we may take it as a fairly safe rule of thumb that differences in "C"$\overline{xx}$(8,1) of 0.5 bits, or greater, are sufficiently improbable on a chance basis as to require explanation on other

Table 3.10 The Structure of Speaker Identity Information in Absolute Voice Ratings of Vocoded Speech

Condition	Speaker Identity Information ["C" $\bar{\bar{x}}x(8,n)$] in Bits							
	PAT I(2,20)[1] Pitch-Mag.	PAT II(8,23) Loud-Rough	PAT III(10,19) Anim.-Rate	PAT IV(12,15) Clar.-Beauty	PAT V(7,9) Normality	SUM	"C" $\bar{\bar{x}}x(8,5)$	R^5
A)Conv. Analog Vocoder with Spect. Flat.(2)	1.90	1.71	2.39	.40	1.17	7.57	5.33 (5)	2.24
B)Conv. Analog Vocoder with Voc. Res. Synth.(2)	1.43	2.08	2.02	.37	1.04	6.94	1.36 (3)	2.58
C)Conv. Analog Vocoder I(2)	1.55	1.30	2.04	.50	.60	5.99	3.27 (3)	2.72
D)Conv. Analog Vocoder II(3)	1.71	1.39	1.97	.00	.62	5.69	3.25 (2)	2.44
E)Conv. Analog Vocoder with Mono. Pitch(3)	.80	1.11	1.44	.00	.34	3.69	1.90 (2)	1.79
F)Conv. Analog(3) Vocoder in Whisper Mode	.90	1.47	1.51	.65	.64	5.17	2.62 (2)	2.55
Aver. for Cond. C and D	1.63	1.35	2.00	.25	.61	5.84	3.26	2.49
Unprocessed Speech(3)	2.07	1.78	2.12	.67	.71	7.35	1.62 (4)	2.73

1. Numbers in parentheses identify items from Speaker Rating Form A which were used to evaluate each PAT.
2. Stimulus materials presented by means of loudspeakers.
3. Stimulus material presented by means of headphones.
4. Dimensionality of listeners response to five PAT's
5. Redundancy = $\sum_{j=1}^{n}$ "C" $\bar{\bar{x}}x(8,j)$ - "C" $\bar{\bar{x}}\bar{\bar{x}}(8,5)$

grounds. Differences of 1.0 bit or greater in "C"$\bar{}_{yy}$(8,5) will be likewise regarded. Where we have occasion to compare averages of "C"$\bar{}_{xx}$(8,1) or "C"$\bar{}_{yy}$(8,5) for two or more conditions, somewhat smaller differences will be regarded as tentatively valid and as meriting some attempt at explanation or interpretation.

Among the more significant trends apparent in Table 3.10 is one which describes the distribution of speaker-identity information across the five PAT's. While there is variation from condition to condition, the first three PAT's consistently carry the bulk of the speaker identity information. PAT V carries a relatively small, but probably significant, amount in most instances, while PAT IV carries from little to no information in all instances. Among other things, therefore, these results raise some questions as to the value of PAT IV for practical purposes of system evaluation. While the results of future research may lead us to discontinue consideration of this factor, a decision on this issue would be premature at this time.

The results for each of the individual vocoders have one or more facets which are of methodological where not of intrinsic significance.

The case of the channel vocoder with spectrum-flattening merits special notice. Taken at face value, these results would indicate that the total speaker-identity information transmitted via the five PAT's is greater than the average for the case of clear speech. In fact, the obtained value of "C"$\bar{}_{yy}$(8,5) is only negligibly smaller than the highest value observed thus far in the course of research with the voice-rating method.

Several considerations lend support to the proposition that the vocoder in question is, in fact, an exceptional vocoder from the standpoint of speaker recognizability. First is the generally-reported "subjective impression" of speaker-recognizability by casual listeners. Second is the objective fact of the ready-recognizability of all voices by the experimenters and other

individuals familiar with the speakers. In addition, the effects of spectrum-flattening upon intelligibility reported in Chapter IV are at best suggestive of a high degree of speaker recognizability, as is the high standing of this vocoder in terms of voice quality. Finally, it is of interest to note that a factor analysis performed to check upon the implicit dimensionality of the five sets of PAT values actually revealed five orthogonal factors - more than have yet emerged in any instance involving the procedures and basic stimulus materials of the first normative study. Barring chance as a significant factor, however, the voice rating results for this vocoder may require further explanation.

To suggest that any form of degradation can actually increase speaker-recognizability would seem to repudiate common sense. Examined more closely, however, the proposition becomes somewhat more tenable. On the hypothesis, for example, that a speaker's natural or average-pitch frequency constitutes one basis for recognition, it is not at all implausible that a pitch extracting vocoder, in particular, could render pitch-frequency more perceptible. Thus, while the acoustic cues to pitch of unprocessed speech are normally rather obscure, once they are accurately evaluated they may well be enhanced by a particular method of speech synthesis. In general, therefore, this and analogous possibilities involving other speech parameters should be borne in mind in assessing the implications for speaker-recognizability of any of the more drastic forms of speech processing found in modern voice communication devices. It is also conceivable that some forms of speech processing may enhance the perceptibility of individual differences in certain characteristics without generally increasing their perceptibility. For example, the increase in "C"$_{\overline{yy}}$ for Factor V, normality, could conceivably be accounted for in terms of the speaker-sensitivity of the system. To the extent that the system operated to degrade certain voices more than others, it could,

in turn, tend to enhance differences in perceived naturalness or normality. Such differences would in turn lead to an increase in the perceptually useful speaker-identity information in degraded speech.

Still other possible explanations of the results for the spectrum-flattened vocoder can be found in terms of perceptual theory — in particular, those aspects concerned with human information-processing behavior. However, an examination of these possibilities will be undertaken at anothe. time in the course of a more general treatment of human information-processing phenomena and their implications for the speaker-recognizability problem.

Subject to verification by additional research, we are, in any case, led to conclude that spectrum flattening enhances the speaker recognizability as well as the intelligibility and voice quality of a conventional 18-channel analog vocoder. An increase of more than two bits of speaker identity information appears attributable to the introduction of spectrum-flattening as a modification of the conventional vocoder.

Several features distinguish the results for the Conventional Vocoder with Vocal Response Synthesizer. Most conspicuous are the high values for total speaker identity information and for information transmitted via PAT's II and V. The physical basis for these results cannot be ascertained at present, though this particular "information pattern" is somewhat suggestive of a high degree of speaker sensitivity — a tendency to affect voices differently with respect to their perceived roughness and "abnormality".

The results for conditions C and D are of interest on several accounts. First is their bearing upon the issues of the reliability and validity of PAT scores. Both conditions involved the same experimental vocoder and stimulus materials. They differed only in method of stimulus presentation. In spite of this difference the results for the two cases are remarkably con-

sistent, both qualitatively and quantitatively. Secondly, these results bear upon the methodological issue of an optimal method of stimulus presentation. In the case of condition C, the speech materials were presented by means of a high-quality loudspeaker; in the case of condition D, Permoflux PDR-8 headphones were employed. While this procedural difference was suggested as a contributing factor to discrepancies between the results of the first and second normative studies, the above results appear to refute such a possibility.

The results for conditions E and F are of special interest in relation to the validity issue in that they provide the occasion for testing specific hypotheses as to the effects of an experimental speech processing upon the information of voice ratings. In the case of condition E, the physical basis of pitch-frequency information is severly degraded. Speech is synthesized with a fixed pitch frequency so as to obscure essentially all individual differences in this parameter. In the case of condition F, both pitch and voicing information are effectively obscured. For both conditions, one would be led inevitably to infer a substantial drop in the value of "C''_{yy}(8,1) for PAT I. From the results in Table 3.10 it can be seen that the hypothesized effect was realized. In both instances there was a substantial reduction in the amount of speaker identity information received via PAT I, while for PAT's II, IV and V no consistent trends of more than negligible degree are apparent. It is of some interest that PAT's II and III, while highly correlated for unprocessed speech, are differentially affected by the treatment in question.

Where the results for unprocessed speech are used as standards for comparison, both of these PAT's appear to sustain a substantial loss in speaker identity information. Where the results for the conventional vocoder are taken as reference values, however, PAT II alone appears to be significantly affected by the loss of pitch frequency information.

A point of special interest is the relationship of the results for conventionally vocoded speech to the results for unprocessed speech. Averages of the results for conditions C and D provide a means of examining this relationship in that they represent the best available data on the effects of a typical vocoder, upon the information, content and structure of voice ratings.

From the results presented in the last two rows of Table 3.10 we are led to conclude that vocoderization does not drastically alter the structure or pattern of the speaker identity information in voice ratings. It appears, in other words, that the effects of vocoderization are not confined primarily to someone or several PAT's. Rather, they are manifested in essentially equal degrees by all of the five PAT's to result in a total loss of approximately 1.3 bits of speaker identity information.

It should be noted, finally, that the effects of various experimental treatments upon the information structure and content of voice ratings are not confined to the values of "C"$\bar{}_{yy}$. Other changes in the structure were revealed by the results of factor analyses performed to check upon the relative orthogonality of PAT values under the various experimental conditions represented in Table 3.10. Shown in parenthesis, next to the value of "C"$\bar{}_{yy}$(8,5) for each

condition, are the number of orthogonal factors revealed by these analyses. It appears from these results that a common consequence of vocoderization is a reduction in the number of independent dimensions of listener response to voices. While this effect stems in part from extreme reduction in the information transmitted via certain PAT's (e.g., as in the case of PAT IV for conditions D and E) it is also attributable in some degree to increases in the correlations among various of the remaining PAT's.

An important implication of the above result would seem to be that: Deprived of stimulus information normally received via a given perceptual channel (i.e., PAT), listeners tend to divert the available "perceptual channel space" to information contained in other stimulus parameters.

An exception to the general trend is found in the case of the conventional vocoder with spectrum flattening. Here the dimensionality of listener response is actually increased relative to the case for clear speech. While this result is possibly an artifact, attributable to the fallibility of the criteria of dimensionality which was employed here, the results are, in any case, consistent with other indications of the superior performance of the "spectrum-flattened" vocoder.

Table 3.11 presents some results of further experiments with vocoded speech. In all of these experiments, the procedures and basic stimulus materials were those employed in the Second Normative Study.

This series of evaluations provided, among other things, an occasion to test for a general relation between speaker identity information structure and the number of vocoder channels employed. However, the results presented in the table reveal no consistent relation. While the total amount

Table 3.11 Structure of Speaker Identity Information in Relative Voice Ratings of Vocoderized Speech

Condition	$\bar{C}_{xx}(m,n)$ in Bits							
	PAT I(4,13)[2] Pitch-Mag.	PAT II(7,10) Loud-Rough	PAT III(8,11) Anim.-Rate	PAT IV(12,15) Clar.-Beauty	PAT V(7,9) Normality	SUM	$\bar{C}_{xx}(m,5)$	R
18-Channel[2] Exper. Vocoder with VRS	1.82	.85	1.35	.32	1.21	5.55	2.80	2.75
11-Channel[2] Exper. Vocoder with VRS	1.91	1.00	1.73	.50	.88	6.05	3.98	2.07
18-Channel Exper. Conv. Vocoder	2.00	.50	1.55	.40	.92	5.27	1.01	1.26
11-Channel Exper. Conv. Vocoder	1.47	.57	1.28	.10	.75	4.15	2.07	2.08
Unprocessed Speech	2.05	1.01	1.62	.71	.80	6.23	1.72	1.51

1. Numbers in parenthesis identify items from Speaker Rating Form IIIA which were used in evaluating each PAT.

2. Because of the unexpected finding that the 11 channel vocoder with VRS transmitted more perceptually useful speaker identity information than its 18 channel counterpart, evaluation of these two vocoders were replicated with a new group of listeners. The trends observed in the first evaluation were unchanged. The informational values presented here were obtained by averaging the results of the two evaluations.

of perceptually significant speaker identity information transmitted by the 18-channel conventional vocoder is approximately twice that transmitted by the 14-channel conventional vocoder, a similar relation does not appear in the case of the vocal response synthesizer. Rather, it was observed in one experiment that substantially more speaker identity information was transmitted by the 14-channel vocoder than by the 18-channel counterpart. A second experiment confirmed the results of the first. Further examination of Table 3.11 reveals that the difference between the two vocoders is not associated with information loss in any one PAT but, rather, represents the cumulation of effects manifested more or less equally by all PAT's. However, additional research will be required to isolate the specific causes of this unpredicted result.

Summary and Recommendations

The results presented in the foregoing section help validate the principles upon which the voice rating method was based and to demonstrate the feasibility of the method for practical purposes of system evaluation. However, several theoretical and practical issues have yet to be fully resolved.

Perhaps most important, on the theoretical side, is the issue of the exhaustiveness with which a five-dimensional voice rating actually characterizes a listener's perception of the distinguishing feature of voice. Two questions arise in this connection.

First is the question of how precisely the method evaluates the amount of speaker identity information transmitted to the typical listener via a single PAT. A number of considerations suggest that the method by no means provides a measure of the typical listener's ultimate capacity for information via a given perceived acoustic trait. Foremost

among these, perhaps, is the unreliability of an overt rating response which is taken as a measure of the underlying perceptual event which evolves it. Listener uncertainty as to the appropriate origin and scale for a rating response unquestionably constitutes a significant source of "noisiness" in the overt characterization of his perceptual experience. Accordingly, the speaker identity information which he transmits to an external observer (or rating form) may thus be substantially less than that which he actually receives and uses in effecting speaker recognition. In addition to these random effects, there is the possibility of systematic variation in the listener's scale and point of subjective "neutrality" (e.g., contextually determined changes in "adaptation level") which, while understood in principle, is difficult to control in practice. However, a fundamental assumption, implicit in the voice rating procedure, is that such extraneous variation at least tends to be uniform over a fairly broad range of experimental speech transmission conditions. Its major implications, therefore, are for the sensitivity rather than the validity of the method. Generally, it should tend to reduce the sensitivity of the method to differences among transmission systems. Techniques for achievement of maximum stability of the listener's reference frame—i.e., adaptation level and response scale—thus merit serious consideration as a subject for future research.

A second question in connection with this issue concerns the true dimensionality of the typical listener's perceptual response to voices. If only on intuitive grounds, we are inclined to reject the possibility that the five dimensions thus far revealed by the voice rating method represent an exhaustive catalog of the basic "ways" by which listeners perceive voices to differ.

Undoubtedly, a great many more perceived characteristics of voices contribute to the recognition process in any one instance. But, while collectively they may carry a relatively large amount of speaker identity information, individually they probably carry only negligible amounts. One reason for this is perhaps a tendency for these minor perceived acoustic traits to be distributed in a highly skewed manner across the population of voices. Thus, while one such trait may contribute significantly to the recognition of a particular speaker from a given sample, it may in general have little or no value for purposes of distinguishing among the remaining speakers in the sample. On the average, therefore, such a trait would carry only a negligible amount of speaker identity information. This is not to deny the possibility that some additional PAT's will emerge once the means are developed for substantially increasing the sensitivity of the voice rating method. It does, however, suggest that the practical significance of such PAT's will tend to be rather small relative to that of the PAT's thus far isolated.

A second major issue relates to the question of the physical-acoustical correlates of the various perceived acoustic traits. Knowledge of these correlates would be of substantial interest in relation to the general theory of voice recognition. It would also serve, however, to enhance the diagnostic uses of voice rating data. Accordingly, it is recommended that future research efforts in the field of speaker recognition involve an intensive search for the physical correlates of perceived acoustic traits.

A third issue concerns the applicability of the multi-dimensional rating method to qualitative variation in speech which is attributable to

factors other than inter-individual differences among speakers. Of particular interest is the problem of system performance with respect to intra-individual differences in speech. This problem is related both to the "quality" problem and to the speaker recognizability problem but it is distinguished from both of them in several important respects. Consider, in particular, its relation to the problem of "quality" as it is treated in Chapter 2.

Here the approach taken to the problem of quality is essentially a negative one. Implicitly, at least, the major basis communication system evaluation from the standpoint of "quality " is freedom from undesirable features such as noise, distraction, hum, and so on. While the fidelity with which certain features of the speech signal are preserved may contribute in various degrees to the value of a gross figure of merit for quality, this latter aspect of system performance is not evaluated explicitly. Under some circumstances, moreover, it may represent an important factor of over-all communications efficiency.

For, in addition to the cues carried in the speech signal concerning the speaker's interest and his identity, there are also perceptible cues to his mood, emotional state, attitude, and even, perhaps, his honesty. The fidelity with which these cues are transmitted represents an aspect of system performance which is amenable to treatment by means of the methods used in evaluating speaker recognizability. For the major issue here is quite analogous to the major issue in the case of voice recognition. This is the issue of the number and nature of the perceptually significant dimensions of intra-individual variation in speech. Thus, the use of multi-dimensional rating techniques as a means of resolving this issue merits

serious consideration as a topic for future research. It is conceivable, in fact, that the same rating forms used in evaluating speaker recognizability may also be used in research on the nature of intra-individual variation in voice quality.

In conclusion, it appears that the major problems yet to be resolved in connection with the voice rating method are of a technological rather than a theoretical nature. All of them, moreover, seem likely to yield to steps towards a general refinement of voice rating methods and procedures.

REFERENCES

1. Asher, J. W., T. D. Hanley, and M. D. Steer, "A Factor Analysis of Twelve Physical Measures of Voice," NAVTRADEVCEN Tech. Report, No. 104-2-48 (1957).

2. Baker, E. J., and E. A. Alluisi, "Information Handling Aspects of Visual and Auditory Form Perception," J. Eng. Psychol. 1, 159 (1962).

3. Brown, R. W., R. A. Leiter, and D. C. Hildum, "Metaphors from Music Criticism," J. Abnorm. Soc. Psychol. 54, 347 (1957).

4. Buck, G. A., "The Conduct of Free Conversation Opinions Tests for Rating Speech Links," Post Office Engineering Dept. Res. Report, No. 20023, London (1959).

5. Buck, G. A., "The Working Reference Telephone Circuit for Speech Link Assessment Studies," Post Office Engineering Dept. Res. Report, No. 20220, London (1960).

6. Buck, G. A., "Rating the Post Office Transmission Standard and Other Transmission Systems with the Reference Speech Link," Post Office Engineering Dept. Res. Report, No. 20561, London (1960).

7. Buck, G. A., and A. F. Beardmore, "Determination of a Rating Scale for Use with the Reference Speech Link in Conversational Assessments," Post Office Engineering Dept. Res. Report, No. 20560, London (1960).

8. Compton, A. J., "Effects of Filtering and Vocal Duration Upon the Identification of Speakers, Aurally," J. Acoust. Soc. Am. 35, 1748 (1963).

9. Dixon, W. J., and Massey, F. J. Jr., "Introduction to Statistical Analysis," (New York, McGraw-Hill, 1957).

10. Eliiot, Lois L. and Tannenbaum , Percy H., "Factor-Structure of Semantic Differential Responses to Visual Forms and Prediction of Factor Scores From Structural Characteristics of the Stimulus Shapes," Amer. J. Psychol. 76, 589 (1963).

11. Fairbanks, G., "Test of Phonemic Differentiation: The Rhyme Test," J. Acoust. Soc. Am. 30, 596 (1958).

12. Guilford, J. P., "Psychometric Methods," (New York, McGraw-Hill, 1954).

13. Gulliksen. H. and Messick, S.(eds.), "Psychological Scaling: Theory and Applications," (New York, Wiley, 1960).

14. Hargreaves, W. A., and J. A. Starkweather, "Recognition of Speaker Identity, Language and Speech 6, 63 (1963).

15. Holmgren, G. L., "Physical and Psychological Correlates of Speaker Recognition," Unpublished Ph.D. dissertation, Texas Christian University, 1964.

16. House, A. S., C. Williams, M. H. L. Hecker, and K. D. Kryter, "Psychoacoustic Speech Tests: A Modified Rhyme Test," Decision Sciences Lab. Tech. Report, No. ESD-TDR-63-403 (1963).

17. Jakobson, Roman, and Halle, Morris, "Fundamentals of Language," S. Gravenhage, Mouton and Co. (1956).

18. Kersta, L. G., "Voice Spectrograms for Unique Personal Identifications," Bell Lab. Rec. 40, 214 (1962).

19. Kersta, L. G., "Voiceprint Identification," J. Acoust. Soc. Am. 34, 725 (1962).

20. Kurtzberg, R. L., M. Alpert, and A. J. Friedhoff, "Identification from Voice: Techniques for the Reduction of Trial-Retrial Variability," J. Acoust. Soc. Am. 35, 1877 (1963).

21. McGee, V. E., "The Determination of a Perceptual Space for the Quality of Filtered Speech," (Educational Testing Service, 1961).

22. Miller, G. A., and P. A. Nicely, "An Analysis of Perceptual Confusions Among Some English Consonants," J. Acoust. Soc. Am. 27, 338 (1955).

23. Munson, W. A., and J. E. Karlin, "Isopreference Method for Evaluating Speech-Transmission Circuits," J. Acoust. Soc. Am. 34, 762 (1962).

24. Ochiai, Y., and T. Fukumura, "Timbre Study of Vocalic Voices," M.F.E., Nagoya University, 5, 253 (1953).

25. Ochiai, Y., and T. Fukumura, "Timbre Studies of Vocalic Voices Viewed from Subjective Phonal Aspect." Part I - preliminary studies on naturalness and articulation qua'ities actually and directly measured with respect to band-eliminating distortion, M.F.E., Nagoya University, 8, 77 (1956).

26. Ochiai, Y., T. Fukumura, and A. Hattori, "Timbre Study of Vocalic Voices Viewed from Subjective Phonal Aspect," Part II (a) - preliminary studies on timbre confusion of phoneme and voice, M.F.E., Nagoya University, 8, 203 (1956).

27. Ochiai, Y., "Phoneme and Voice Identification Studies Using Japanese Vowels," Language and Speech 2, 132 (1959).

28. Ostwald, P. F., "Visual Denotation of Human Sounds," Arch. Gen. Psychiat. 3, 25/117-29/121 (1960).

29. Philco Corporation, Final Report, Contract No. AF19(628)-586, (1965).

30. Pollack, I., J. M. Pickett, and W. H. Sumby, "On the Identification of Speakers by Voice," J. Acoust. Soc. Am. 26, 403 (1954).

31. Pruzansky, S., "Pattern-Matching Procedure for Automatic Talker Recognition," J. Acoust. Soc. Am. 35, 254 (1963).

32. Pruzansky, S., and M. V. Mathews, "Talker-Recognition Procedure Based on Analysis of Variance," J. Acoust. Soc. Am. 35, 1877 (1963).

33. Rothauser, E. H., "Modified Isopreference Method for Audio-Quality Measurements," J. Acoust. Soc. Am. 35, 1899 (1963) (Abs.)

34. Shearme, J. N., and J. N. Holmes, "An Experiment Concerning the Recognition of Voices," Post Office Engineering Dept. Res. Report, No. 20513, London (1959).

35. Shearme, J. N., and J. N. Holmes, "An Experiment Concerning the Recognition of Voices," Language and Speech 2, 123 (1959).

36. Smith, J. E., "Decision-Theoretic Speaker Recognizer, " J. Acoust. Soc. Am. 34, 1988 (1962).

37. Solomon, L. N., "Semantic Approach to the Perception of Complex Sounds," J. Acoust. Soc. Am. 30, 421-425 (1958).

38. Solomon, L. W. "Search for Physical Correlates to Psychological Dimensions of Sounds," J. Acoust. Soc. Am. 31, 491-497 (1959).

39. Starkweather, J. A., and W. A. Hargreaves, "Speaker Identification and Vocal Variability," J. Acoust. Soc. Am. 34, 1978 (1962).

40. Steer, M. D., T. D. Hanley, and R. C. Bilger, "A Further Investigation of the Relationships Between Voice Variables and Speech Intelligibility in High-Level Noise," SPECDEVCEN Tech. Report, No. 104-2-26 (1955).

41. Stevens, K. N., "Simplified Nonsense-Syllable Tests for Analytic Evaluation of Speech Transmission Systems," J. Acoust. Soc. Am. 34, 729 (1962) (Abs.).

42. Stuckey, C. W., "Investigation of the Precision of an Articulation-Testing Program," J. Acoust. Soc. Am. 35, 1782 (1963).

43. Tarnoczy, T. H.,"Determination of the Speech Spectrum Through Measurements of Superposed Samples," J. Acoust. Soc. Am. 28, 1270 (1956).

44. Torgerson, W. S., Theory and Methods of Scaling (New York, Wiley, 1958).

45. Voiers, W. D., "Perceptual Criteria of Speaker Identity," J. Acoust. Soc. Am. 33, 1677 (1961).

46. Voiers, W. D., "The Perceptual Bases of Speaker Identity," J. Acoust. Soc. Am. 36, 1065 (1964).

APPENDIX I

LISTENERS, SPEAKERS AND APPARATUS

LISTENERS, SPEAKERS AND APPARATUS

LISTENERS: Several different groups of listeners were used in the course of the program. For purposes of various exploratory studies, several groups of untrained and unpracticed male college students from Tufts University were used. In subsequent discussions, any one such group is designated as "Group X".

One large-scale study of voice recognition was conducted at Brandeis University. It involved four groups of eight students, four male and four female; these groups will subsequently be referred to as Groups 1A, 1B, 1C and 1D.

For the main body of research, four groups of eleven normal-hearing, male students at Tufts University were used on a weekly basis over a period of 12 weeks, for experiments conducted at the Experimental Social Psychology Laboratories of Tufts University. These are designated as Groups 2A, 2B, 2C and 2D. Ordinarily, data were evaluated only for eight listeners in each group. Three alternates performed primarily to maintain levels of experience in the testing situation. For all experimental studies of intelligibility and quality conducted at Tufts, eight listeners were housed, two to a booth, in four sound-treated booths. The remaining listeners were housed in an adjacent classroom where they also received all of the stimulus material via headphones.

During the final weeks of the program, a group of ten normal hearing, male college freshmen participated on a weekly basis in a series of experiments conducted at the Sperry Rand Research Center. These subjects are referred to as Group 3 in subsequent sections of this report.

SPEAKERS: A pool of 26 male employees of the Sperry Rand Research Center, selected for general American dialectal background, provided the recorded speech materials used in the program. One of these (RD) was experienced as a radio announcer and made recordings of Diagnostic Rhyme Test materials which were used extensively throughout the program. Another speaker (RC) was selected on the basis of results from a voice rating study which indicated his voice to be most nearly "neutral" in terms of five perceived acoustic traits. This speaker recorded materials which were used routinely in system evaluations made with the Diagnostic Rhyme Test. He also served as the announcer and "reference voice" in a series of relative voice-rating studies and as one of five speakers in a series of speech-quality evaluation studies.

Twenty-four other speakers were used in the course of several voice-recognition studies. Six of these, selected on the basis of their perceived voice characteristics (high pitch, low pitch, high loudness-animation, low loudness-animation, high clarity and low clarity) were used in various experiments on speech intelligibility and on speech quality.

EQUIPMENT: With minor variations in certain instances, the basic complement of experimental equipment used in experimental studies of intelligibility and quality consisted of the following:

Crown SS822 Tape Recorder

Scott Type 246 Dual Channel Audio Amplifier

Permoflux PDR-8 matched headphones

2 Krohn-Hite Model 330M variable band-pass filters

Maico audiometer, Model MA-12

A specially constructed reference standard system for stimulus presentation was used on occasion during the final stages of the program.

Speech materials used in all three phases of the program were recorded at the Sperry Rand Research Center. Recording equipment and facilities included the following:

Electro voice Model 300 Dynamic Microphone

Crown SS1433 Tape Recorder

A Sound-treated (26 dB) room (6' x 6' x 7') manufactured by Silence, Inc.

In all speech recordings the speaker was seated in a chair in the center of the sound-treated room, with head held in a fixed position 40 cm from the microphone. The speech material to be recorded was placed on a stand in front of the speaker so that it could be seen easily without changing the position of the head. A small light, adjusted to flash at the appropriate rate, was placed next to the speech material to be recorded. This was used by the speaker to time his utterances. All recording was done on 1/2" tape, at a speed of 7-1/2 OPS, using a recorder situated outside the sound-treated room. Recording level was adjusted so that the "average-peak" VU readings were -3 dB.

The resulting recordings were then edited and copied onto 1/4" tape and it was the copied recordings which were used as "working masters".

APPENDIX II

SAMPLES OF BIOGRAPHICAL DATA SHEET AND SENTENCE LISTS USED IN VOICE RATING STUDIES AND IN SPEECH QUALITY

Biographical Data Sheet Administered to Listeners and Speakers

Code_____

BIOGRAPHICAL SKETCH

Date_____

Name_____ Marital Status_____

Employer_____ Occupation_____

Date of Birth_____ Height_____ Weight_____

Present Address_____
(Street) (City) (State) (Tel. No.)

Place of Birth_____
(City) (State)

Cities Lived Prior to 16th Birthday

City	State	Age

Education Received

	City	State
High School		
College		

Highest Degree_____ Major_____

Approximate Standing at Graduation (at highest level)

Upper 1/3 _____ Middle 1/3 _____ Lower 1/3 _____

Parents

Place of Birth:	City	State	Highest Grade Completed	Occupation
Father:				
Mother:				

What, if any, foreign languages were spoken by members of your family as a child?

_____

Was any particular national group(s) predominant in communities where you were raised?

Community	National Group

What foreign languages do you speak?

Language							
_____	Poorly	____	____	____	____	____	____ Well
_____	Poorly	____	____	____	____	____	____ Well
_____	Poorly	____	____	____	____	____	____ Well

Where have you spent the greatest part of your adult life?

_____

Have you had any dramatic or public speaking training? _____ Experience_____

_____

Stimulus Materials Used for Voice Rating Studies*

1. THE WATER'S TOO COLD FOR SWIMMING.
2. HERE ARE YOUR SHOES THIS TIME.
3. YOU SHOULD COME HERE WHEN I CALL.
4. DON'T USE UP ALL THE LETTER PAPER.
5. THOSE PEOPLE OUGHT TO SEE A DOCTOR.
6. THE WINDOWS ARE SO DIRTY I CAN'T SEE.
7. DON'T LET THE DOG OUT OF THE HOUSE.
8. WAIT FOR ME OVER IN THE PARK.
9. IF YOU WANT ANYTHING, JUST CALL.
10. PUT THAT BIG BOX UNDER THE BED.
11. I CAN'T GO WITH YOU THIS MONTH.
12. YOU CAN CATCH THE BUS ACROSS THERE.
13. TELL HER THE NEWS ON THE PHONE.
14. I'LL CATCH UP WITH YOU LATER.
15. I'LL THINK IT OVER AND CALL HER.
16. I DON'T WANT TO GO TO THE MOVIES.
17. IF YOUR TOOTH HURTS SEE A DENTIST.
18. PUT THAT COOKIE BACK IN THE BOX.
19. HE OUGHT TO STOP FOOLING AROUND.
20. TONIGHT THAT MUCH TIME'S UP.
21. I DON'T KNOW HOW TO SPELL HIS NAME.
22. YOU CAN FIND IT DOWN THE STREET.
23. WALKING'S MY FAVORITE EXERCISE.
24. HERE'S A NICE QUIET PLACE TO REST.

*Sentences 1-24 were used in the first normative study and associated experimental studies; sentences 1-16 were used in the second normative study and associated experimental studies.

List of Sentences Used in Tests of Speech Quality.
(Formally-Trained Speaker Recorded These Sentences)

1. Don't try to finish them before Tuesday.
2. He knows how to paddle a canoe.
3. There was oil spilled all over the road.
4. I think I'll eat in the cafeteria tomorrow.
5. The United Charity Fund exceeded its goal.
6. He would like to try again today.
7. The current rate is only three per cent.
8. I think I'll go down town this afternoon.
9. A drill press is a useful tool to have.
10. You have to judge the time very accurately.
11. Tap on the door and then go in.
12. It was a good thing for him to do.
13. The traffic gets very heavy after five.
14. Leave your package there on the bench.
15. His best score was over two hundred.
16. It's hard to tell who's the best man for the job.
17. The arm of the chair was worn thin.
18. You should be able to do it in a couple of hours.
19. His car can easily pass the train.
20. Don't pay any attention to the check marks.
21. I'm afraid the rain won't stop before noon.
22. It was down to ten below zero last night.
23. The whole town was talking about it.
24. They were still tied in the fourteenth inning.
25. It will be a pleasure to talk in your town.
26. It's hard to tell who's the best man for the job.
27. Don't pay any attention to the check marks.
28. The current rate is only three per cent.

An Answer Sheet and Lists of Sentences Used in Collecting the Preference Data

Listener_____

Code_____

Practice Set

		1	2
1.	I hate those lofty quotations	____	____
2.	Taste is the feminine of genius	____	____
3.	Men desire to be immortal	____	____
4.	Charity must begin at home	____	____
5.	Life is an incurable disease	____	____
6.	Persuasion hung upon his lips	____	____
7.	Cowardly dogs bark the loudest	____	____
8.	Marriage is a desperate thing	____	____
9.	He established law and justice	____	____
10.	Ticker tape is not spaghetti	____	____

Set 1

		1	2
1.	Men desire to be immortal	____	____
2.	Charity must begin at home	____	____
3.	He established law and justice	____	____
4.	Ticker tape is not spaghetti	____	____
5.	Life is an incurable disease	____	____
6.	Persuasion hung upon his lips	____	____
7.	Cowardly dogs bark the loudest	____	____
8.	Marriage is a desperate thing	____	____
9.	I hate those lofty quotations	____	____
10.	Taste is the feminine of genius	____	____

Listener_____

Code_____

Set 2

1. Life disease ___ ___
2. Persuasion. .lips ___ ___
3. Cowardly.. loudest ___ ___
4. Marriage.. thing ___ ___
5. Men. immortal ___ ___
6. Charity.. home ___ ___
7. He.. justice ___ ___
8. Ticker spaghetti ___ ___
9. I .quotations ___ ___
10. Taste . genius ___ ___

Set 3

1. I quotations ___ ___
2 Taste genius ___ ___
3. Life disease ___ ___
4 Persuasion .lips ___ ___
5. Men immortal ___ ___
6. Charity home ___ ___
7. He. justice ___ ___
8 Ticker spaghetti ___ ___
9 Cowardly .loudest ___ ___
10. Marriage thing ___ ___

Set 4

1 He justice ___ ___
2 Ticker spaghetti ___ ___
3 Life disease ___ ___
4 Persuasion lips ___ ___
5 Cowardly loudest ___ ___
6 Marriage thing ___ ___
7 I quotations ___ ___
8 Taste genius ___ ___
9 Men immortal ___ ___
10 Charity. home ___ ___

Set 5

1. Cowardly...loudest ___ ___
2. Marriage...thing ___ ___
3. Life .disease ___ ___
4 Persuasion.. lips ___ ___
5. Men immortal ___ ___
6. Charity...home ___ ___
7 He ..justice ___ ___
8 Ticke .. spaghetti ___ ___
9. I...quotations ___ ___
10. Taste...genius ___ ___

Set 6

1 Cowardly loudest ___ ___
2. Marriage thing ___ ___
3. I . quotations ___ ___
4 Taste genius ___ ___
5. Men immortal ___ ___
6. Charity. .home ___ ___
7. Life ..disease ___ ___
8. Persuasion .lips ___ ___
9. He . justice ___ ___
10. Ticker spaghetti ___ ___

Set 7

1 Cowardly loudest ___ ___
2 Marriage thing ___ ___
3 He justice ___ ___
4 Ticker spaghetti ___ ___
5 Life disease ___ ___
6 Persuasion. lips ___ ___
7 I quotations ___ ___
8 Taste genius ___ ___
9 Men .immortal ___ ___
10. Charity home ___ ___

Set 8

1. He...justice ___ ___
2 Ticker ..spaghetti ___ ___
3 Men ..immortal ___ ___
4. Charity...home ___ ___
5. I...quotations ___ ___
6. Taste. .genius ___ ___
7. Cowardly...loudest ___ ___
8. Marriage...thing ___ ___
9. Life . disease ___ ___
10. Persuasion...lips ___ ___

Set 9

1 I...quotations ___ ___
2. Taste genius ___ ___
3. Men. immortal ___ ___
4 Charity home ___ ___
5. Life disease ___ ___
6. Persuasion .lips ___ ___
7. Cowardly...loudest ___ ___
8. Marriage ..thing ___ ___
9 He justice ___ ___
10. Ticker spaghetti ___ ___

Set 10

1. I quotations ___ ___
2. Taste .genius ___ ___
3 Men immortal ___ ___
4. Charity home ___ ___
5 Life disease ___ ___
6. Persuasion. lips ___ ___
7 Cowardly .loudest ___ ___
8 Marriage ..thing ___ ___
9 He justice ___ ___
10 Ticker spaghetti ___ ___

APPENDIX III

SUMMARIES OF MAJOR EXPERIMENTAL STUDIES

Summary of Experimental Study No. I-1 Date: 11/9/64

Title: Effects of band limited noise upon Diagnostic Rhyme Test scores.

Responsible Scientist(s): WV, MC

Purpose: To determine the performance characteristics of the DRT under various S/N conditions.

Methods & Materials:

Subjects: 40 male employees of SRRC (5 groups of 8). None had previous exposure to the test.
Location: SRRC

Stimulus Materials: DRT - experimental form

Stimulus Conditions: Speech mixed with filtered white noise (60-7500 Hz) at S/N ratios of -12dB, -6dB, 0 dB, +6dB, +12dB.

Equipment: Crown tape recorder
Scott amplifier
noise generator
PDR-8 matched earphones
Krohn-Hite Filter

Experimental Design:

Each of 5 groups listened to one of the 5 S/N conditions. 4 subjects of each group listened to tapes of words with feature present first, and 4 of each group listened to feature absent first.

Results & Discussion : See following figure.

Summary & Conclusions:

DRT total score has a gain function: $\frac{\Delta DRT}{\Delta \frac{S}{N}} = 2\%$

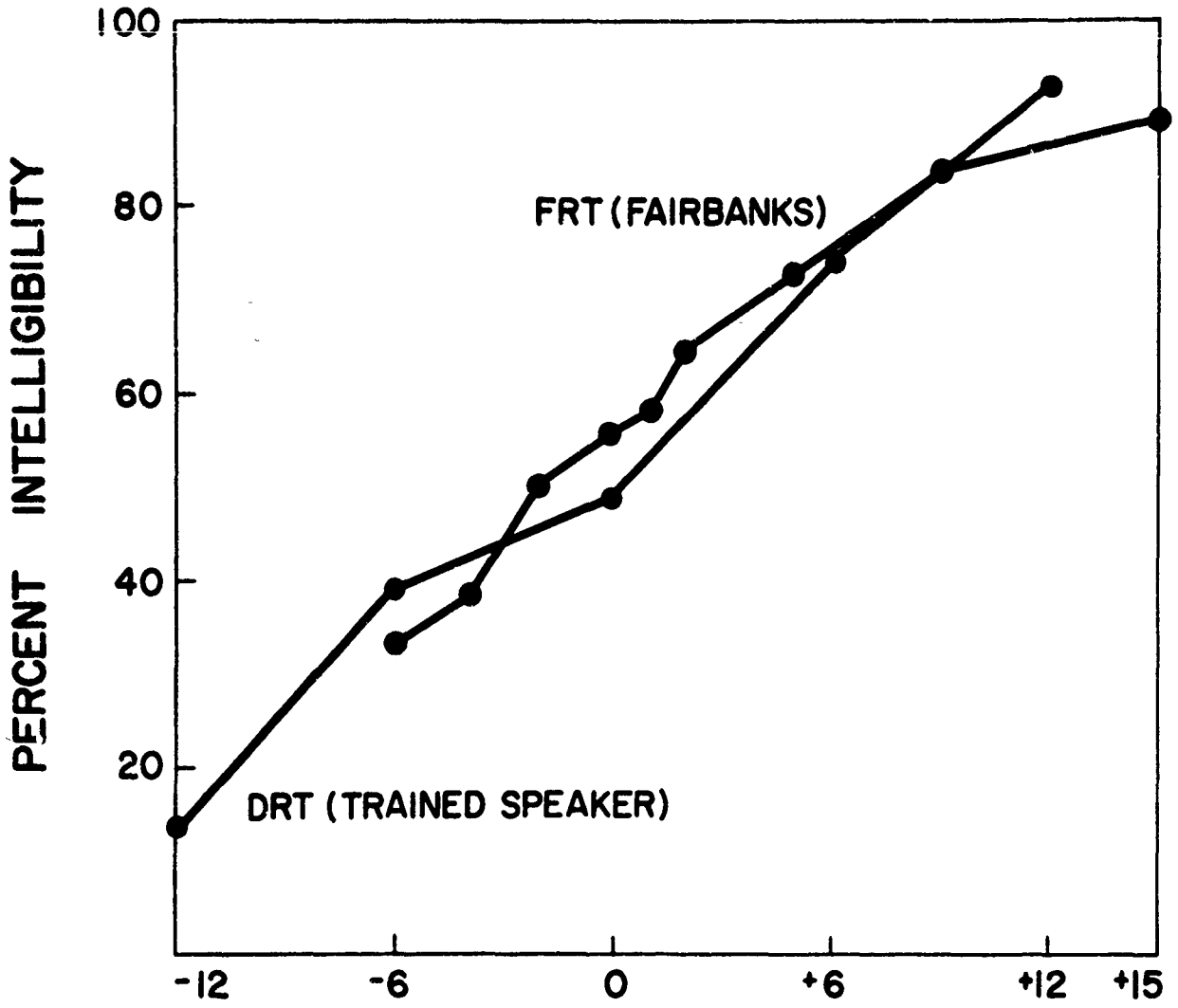


FIG. I-1a Effects of noise upon intelligibility test scores.

Summary of Experimental Study No. I-2 Date. 3/22/65

Title: Comparative evaluation of two rhyme tests.

Responsible Scientist(s): MC,WV

Purpose: To determine the performance characteristics of the DRT and FRT under similar speech-to-noise conditions.

Methods & Materials:

Subjects: 32 male university students (groups 2A, 2B, 2C, 2D).

Location: Tufts University

Stimulus Materials: DRT (random form neutral speaker) - 16 lists.
FRT (neutral speaker) 4 random list orders.

Stimulus Conditions: Speech mixed with filtered white noise (200-4 Hz) at S/N levels of -9 dB, 0 dB, +9 dB, and +18 dB.

Equipment: Crown tape recorder
Scott amplifier
noise generator
PDR-8 matched earphones
Krohn-Hite Filter

Experimental Design: For each S/N ratio 4 DRT lists and 1 FRT list was used. Each group listened to all 4 S/N conditions, hearing different lists for each condition. The order of presentation was as follows:

I	1a	2b	3c	4d	a = +18 dB S/N
II	4c	3d	2a	1b	b = +9 dB S/N
Group III	3b	4a	1d	2c	c = 0 dB S/N
IV	2d	1c	4b	3a	d = -9 dB S/N
	1	2	3	4	
	Order of Presentation				

1 = DRT list 1-4
FRT list 1
2 = DRT list 5-8
FRT list 2
3 = DRT list 9-12
FRT list 3
4 = DRT list 13-16
FRT list 1

Results & Discussion:

Results indicate that the FRT yields a somewhat higher score that the DRT under identical speech-to-noise conditions.

See following Figure.

Summary & Conclusions:

DRT scores are differentially effected by masking noise. Reliability of DRT scores decreases with value of score over the range from 60-100% intelligibility.

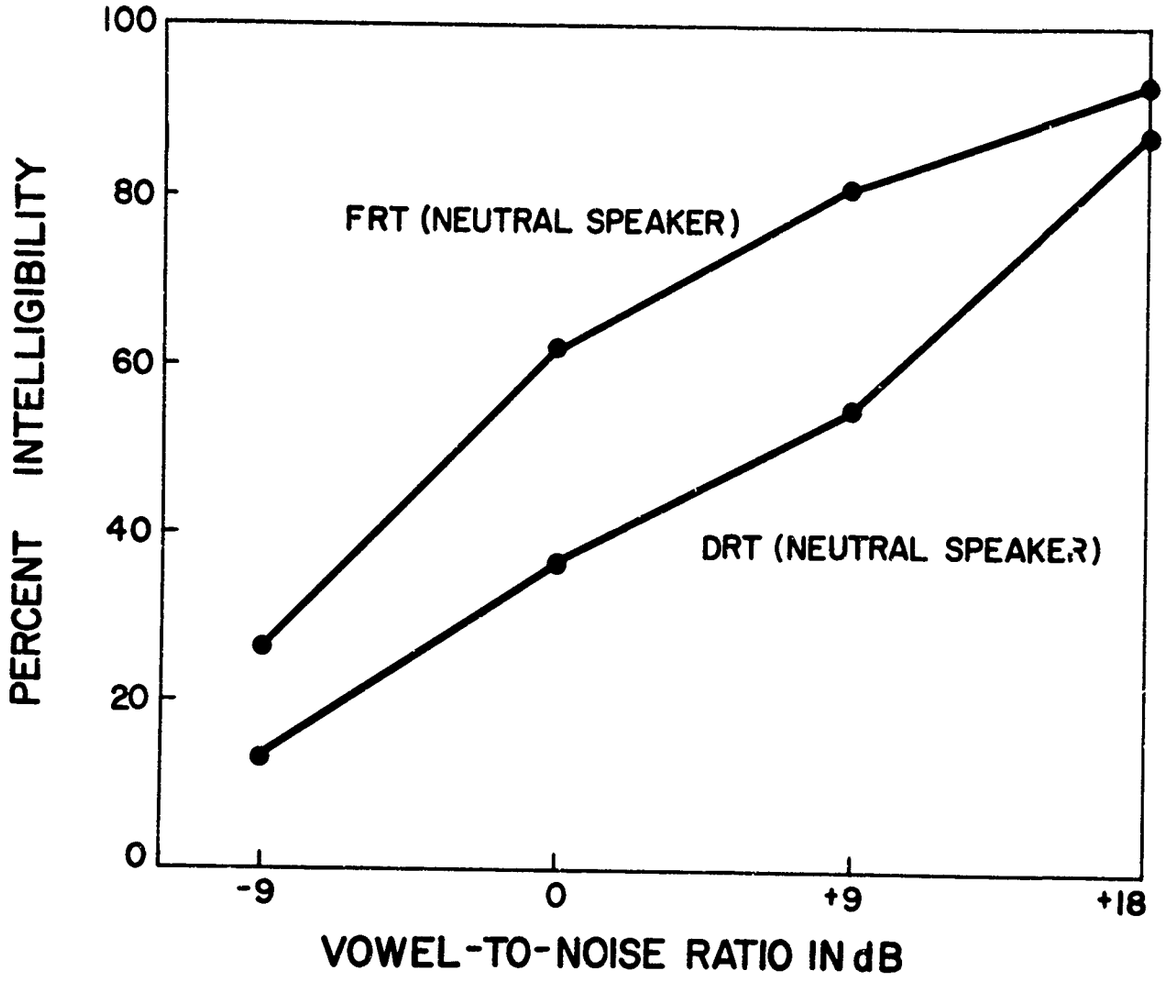


FIG. I-2a A comparison of Diagnostic Rhyme Test and Fairbanks Rhyme Test scores under various noise conditions.

Summary of Experimental Study No. I-3 Date: 1/15/65

Title: Effects of stimulus-presentation rate upon diagnostic intelligibility scores.

Responsible Scientist(s): MC

Purpose: To determine the optimum rate of stimulus presentation for intelligibility testing.

Methods & Materials:

Subjects: 20 male university students (Group X)

Location: Tufts University

Stimulus Materials: DRT (random form, trained speaker) 5 lists
FRT (trained speaker) 5 lists

Stimulus Conditions: Vocoded and unprocessed speech presented at stimulus rates of one word every 2.8 sec., 2.0 sec., 1.4 sec., 1.0 sec., and 0.7 sec.

Equipment: Crown recorder - 2 channel
Eico amplifier (Ch. I-unprocessed tape)· Krohn-Hite filter 100-5K
Scott amplifier (Ch. II - vocoderized tape)
8 sets of matched PDR-8 earphones

Experimental Design:

Each group heard both the DRT and FRT at all 5 stimulus rates in different orders. Half of each group (2 subjects) heard vocoderized tapes (Ch. II) and half listened to unprocessed tapes (Ch. I).

	1	2.8	2.0	1.4	1.0	0.7	
	2	2.0	1.4	1.0	0.7	2.8	
Order of Presentation	3	1.4	1.0	0.7	2.8	2.0	
	4	1.0	0.7	2.8	2.0	1.1	
	5	0.7	2.8	2.0	1.4	1.0	V U
		1	2	3	4	5	Condition

Results & Discussion:

Results indicate that the rate of one word every 1.33 seconds yields the smallest standard errors of the mean and also the highest intelligibility score for both the DRT and the FRT. This occurs for vocoded speech and for unprocessed speech.

Summary & Conclusions:

The DRT and FRT, vocoderized and unprocessed, were presented to listeners at 5 stimulus presentation rates. On the basis of resulting intelligibility scores and standard errors of the mean, the rate of one word every 1.4 seconds was chosen as the optimum rate of stimulus presentation for intelligibility testing.

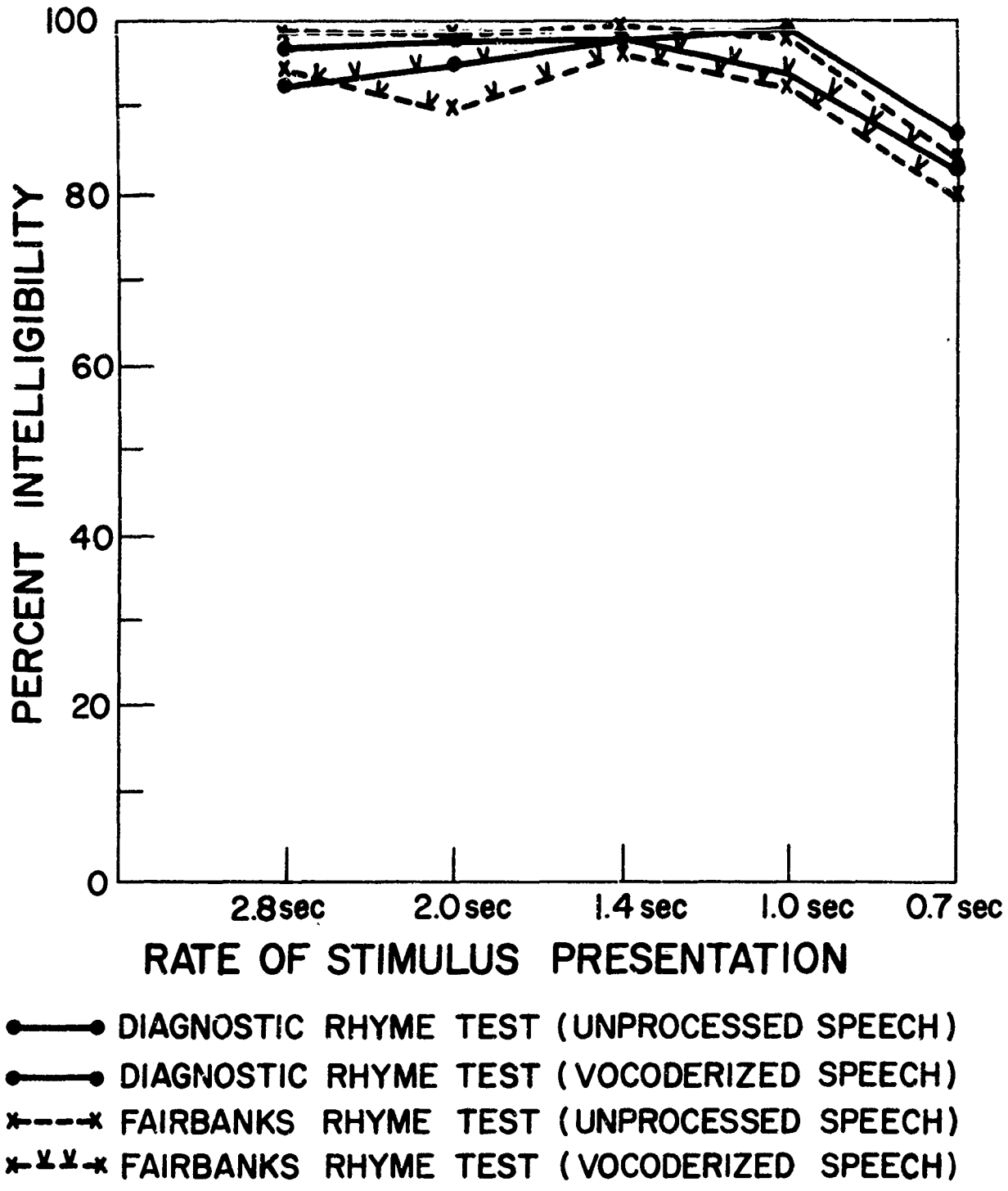


FIG. I-3a Effects of stimulus presentation rate upon speech intelligibility.

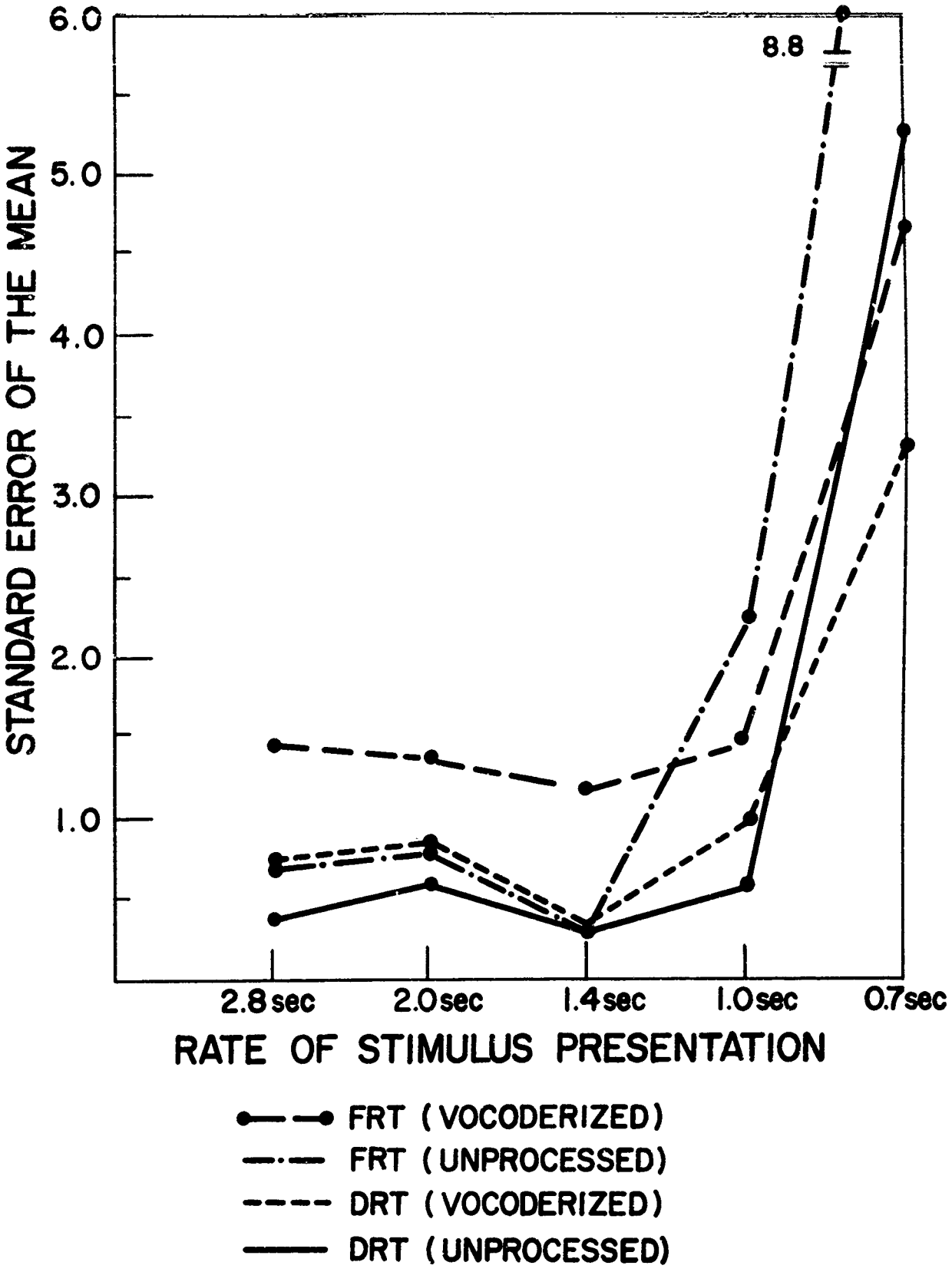


FIG. I-3b Effects of stimulus presentation rate upon standard errors of rhyme test scores.

Summary of Experimental Study No. I-4 Date: 3/11/65

Title: Further research on effects of stimulus-presentation rate on diagnostic intelligibility scores.

Responsible Scientist(s): MC, WV

Purpose: To determine the optimum rate of stimulus-presentation for intelligibility testing of vocoders.

Methods & Materials:

Subjects: 16 male university students (Group X)

Location: Tufts University

Stimulus Materials: DRT (random form, neutral speaker) 4 lists
FRT (neutral speaker) 4 lists

Stimulus Conditions: Speech processed by 4 experimental vocoders and presented at stimulus rates of one word every 2.0 sec., 1.66 sec., 1.33 sec, and 1.0 sec.

Equipment: Crown Tape Recorder PDR-8 Matched Earphones
Scott Amplifier
Noise Generator

Experimental Design: Each of 4 groups listened to tapes of the DRT and FRT as processed by one of 4 vocoders. Order of rate of stimulus presentation was 2.0 sec., 1.66 sec., 1.33 sec., and 1.0 sec. for all groups.

Results & Discussion: See below.

Summary & Conclusions: The DRT and FRT were recorded using 4 stimulus presentation rates. The tapes were then processed by 4 experimental vocoders and each vocoder tape was presented to a group of 8 listeners. Resulting intelligibility scores indicate that when words are presented at the rate of 1 every 1.33 seconds, there is no adverse effect on either the intelligibility score or the standard error score.

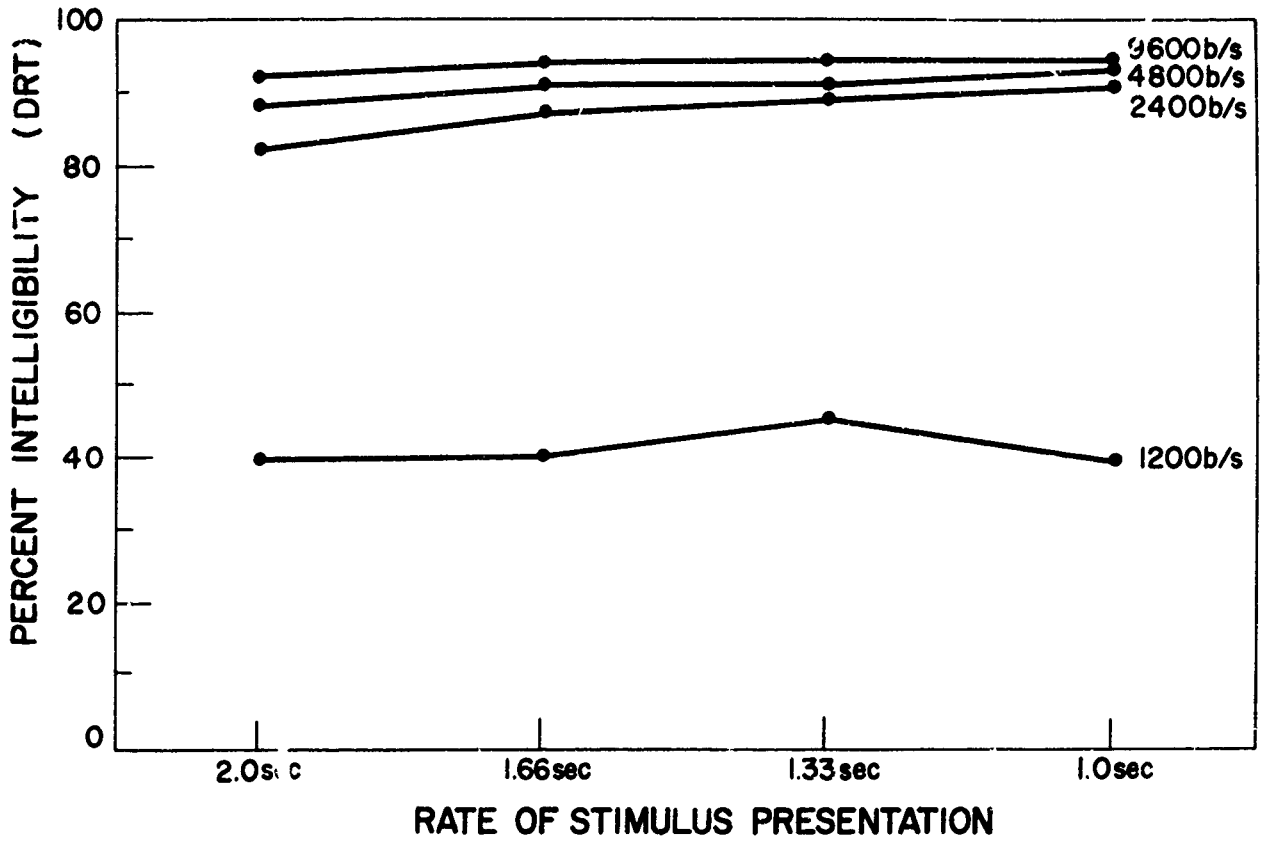


FIG. I-4a Effects of stimulus presentation rate upon Diagnostic Rhyme Test total score for four experimental vocoders.

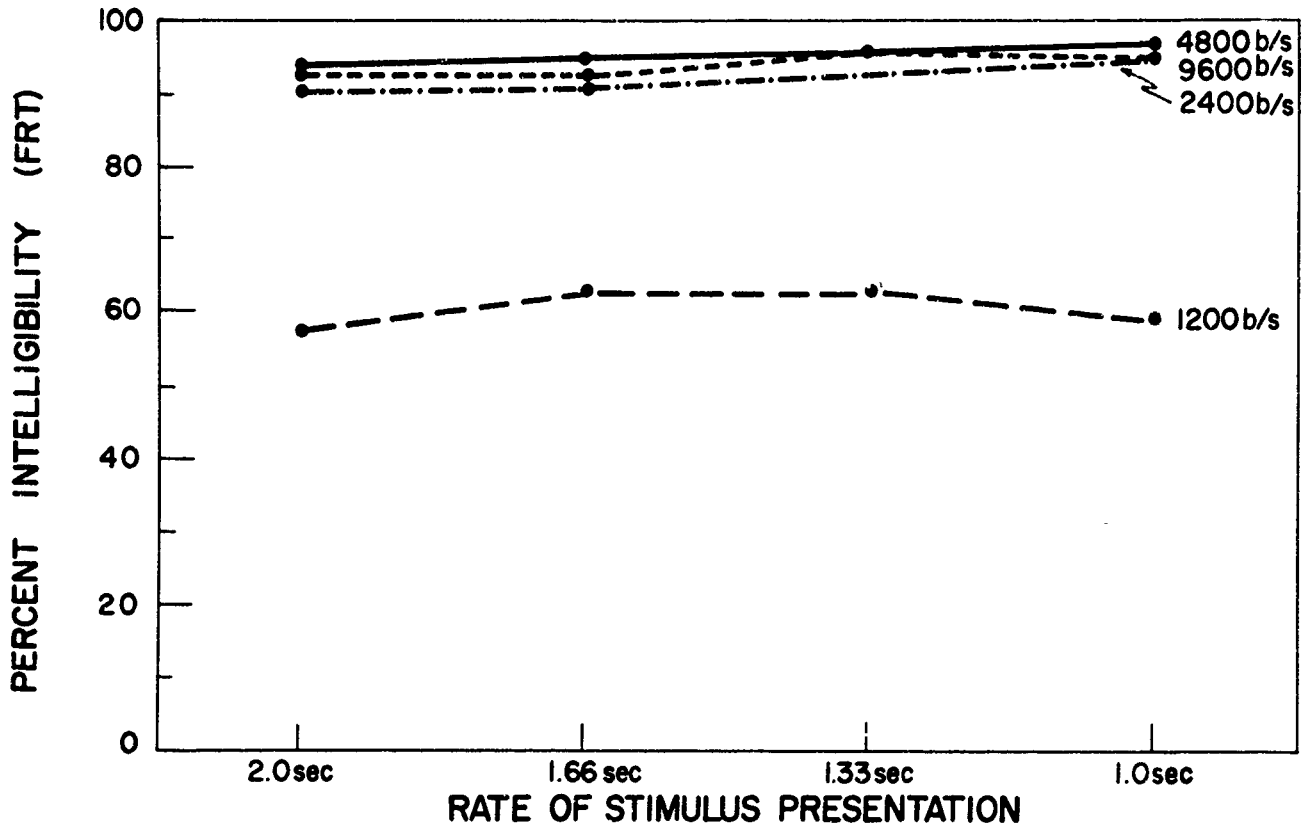


FIG. I-4b Effects of stimulus presentation rate upon Fairbanks Rhyme Test score for four experimental vocoders.

Summary of Experimental Study No. I-5 Date: 4/26/65

Title: Effects of multiple vocoderization on DRT scores.

Responsible Scientist(s): MC,WV

Purpose: To determine the effects on intelligibility of speech which has been processed by a vocoder 1, 2, and 3 times.

Methods & Materials:

Subjects: 32 male university students (Groups 2A, 2B, 2C and 2D)

Location: Tufts University

Stimulus Materials: DRT (random list, neutral speaker) 4 lists

Stimulus Conditions: Speech processed by each of 4 experimental vocoders 3 times, 2 times, and one time, and unprocessed speech.

Equipment: Crown Tape Recorder
Scott Amplifier
Noise Generator
PDR-8 Matched Earphones

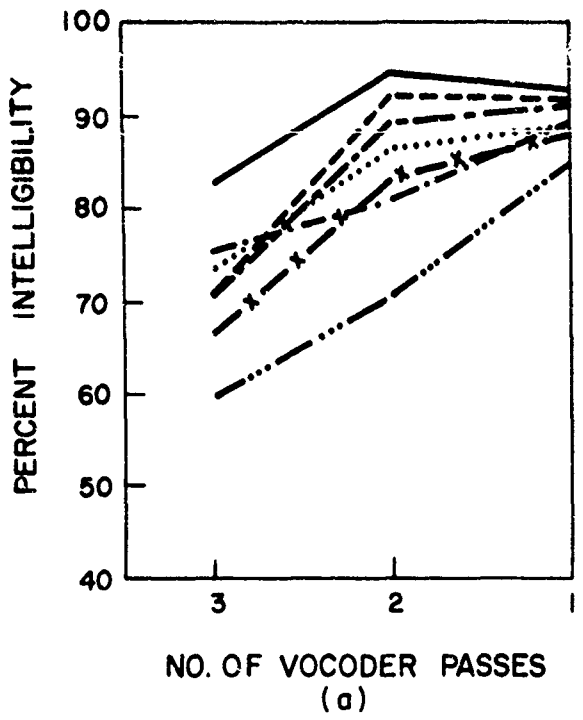
Experimental Design: Each group listened to recordings processed 3 times, 2 times, and one time by the 4 experimental vocoders, and also to unprocessed speech.

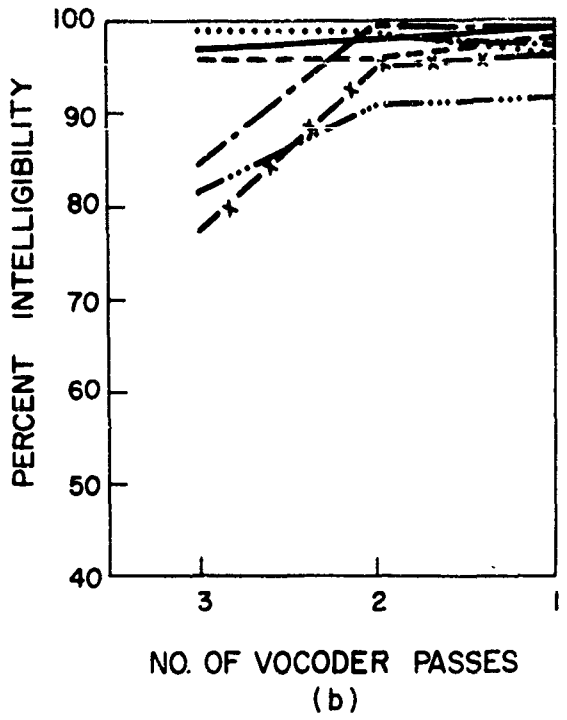
Results & Discussion:

See attached figure

Summary & Conclusions:

Individiual diagnostic scores are differently affected by multiple vocoderization.





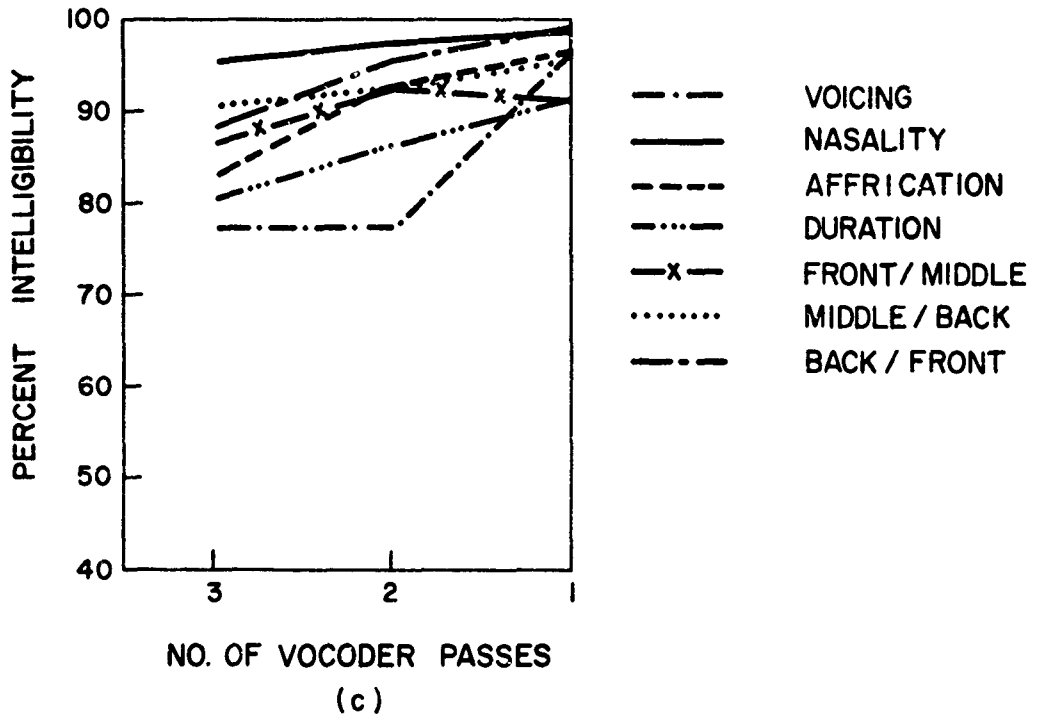


FIG. I-5a Effects of multiple vocoderization upon Diagnostic Rhyme Test scores.
(a) experimental pitch-excited vocoder
(b) experimental voice-excited vocoder
(c) experimental pitch-excited vocoder

Summary of Experimental Study No. I-6 Date: 6/12/65

Title: Speaker effects upon the intelligibility of vocoded speech.

Responsible Scientist(s): WV, MC

Purpose: To determine the effects on intelligibility of vocoded speech of several selected speakers.

Methods & Materials:

Subjects: 32 male university students (Groups 2A, 2B, 2C, 2D)

Location: Tufts University

Stimulus Materials: DRT-(random form, 8 speakers) 8 lists

Stimulus Conditions: Speech of 8 speakers as processed by 4 experimental vocoders.

Equipment: Crown Tape Recorder
Scott Amplifier
Noise Generator
PDR-8 Matched Earphones

Experimental Design: Each of 4 groups listened to 8 DRT lists, 1 list as spoken by 8 selected speakers and processed by one of 4 experimental vocoders

Results & Discussion: Results indicate a fairly systematic speaker influence upon intelligibility scores. A coefficient of concordance was computed in order to evaluate the consistency of inter-speaker differences across vocoders. The obtained value of .76 $P \leq .13$, is strongly suggestive of differences in the inherent intelligibility of individual voices as transmitted by vocoders.

Summary & Conclusions:

Speakers differ significantly in inherent intelligibility as measured by the DRT. Speaker effects are interactive with vocoder effects.

Summary of Experimental Study No. I-7 Date: 8/11/65

Title:Effects of frequency pass band upon DRT scores.

Responsible Scientist(s): WV; MC

Purpose: To determine the effects of frequency filtering on DRT scores.

Methods & Materials:

Subjects: 8 university students (Group 3)

Location: SRRC

Stimulus Materials:DRT (random form, trained speaker).

Stimulus Conditions: Stimulus materials were high passed at 3200 Hz, 2250 Hz, 1590 Hz, and 1125 Hz, and were low passed at the same frequencies. Listening level was adjusted to approximately 85 dB S/N for non-filtered speech and remained constant for all conditions.

Equipment:

Crown tape recorder	2 Krohn-Hite band pass filters
Scott amplifier	PDR-8 matched earphones

Experimental Design:

Subjects listened to the DRT under each of the 8 conditions of frequency filtering.

Results & Discussion

Due to a malfunction of equipment, results are not noteworthy.

Summary & Conclusions:

N/A

Summary of Experimental Study No. I-8 Date: 5/3/65

Title: Effects of conventional, monotone and whisper vocoderization on DRT scores.

Responsible Scientist(s): WV, MC

Purpose: To obtain DRT scores for 3 modes of a conventional 18 channel vocoder.

Methods & Materials:

Subjects: 32 male university students (Groups 2A, 2B, 2C, 2D)

Location: Tufts University

Stimulus Materials: DRT (experimental form, trained speaker)

Stimulus Conditions: Speech materials processed by an 18 channel analog vocoder operating as a conventional vocoder, a monotone vocoder, and as a whispering vocoder.

Equipment: Crown Tape Recorder
Scott Amplifier
Noise Generator
PDR-8 Matched Earphones

Experimental Design: Each of 4 groups listened to the experimental form of the DRT. Three groups were used to obtain scores for the 3 conditions mentioned above. The fourth group listened to an unprocessed recording of the DRT.

Results & Discussion: The 3 modes of the vocoder are very nearly alike with respect to all features except voicing. The whisper vocoder yields a score of 72% for transmission of voicing cues, while the monotone and conventional modes yield scores of 88 and 90% respectively.

Summary & Conclusions: The DRT was used to evaluate an 18-channel analog vocoder operating as a conventional vocoder, a monotone vocoder, and as a whispering vocoder.

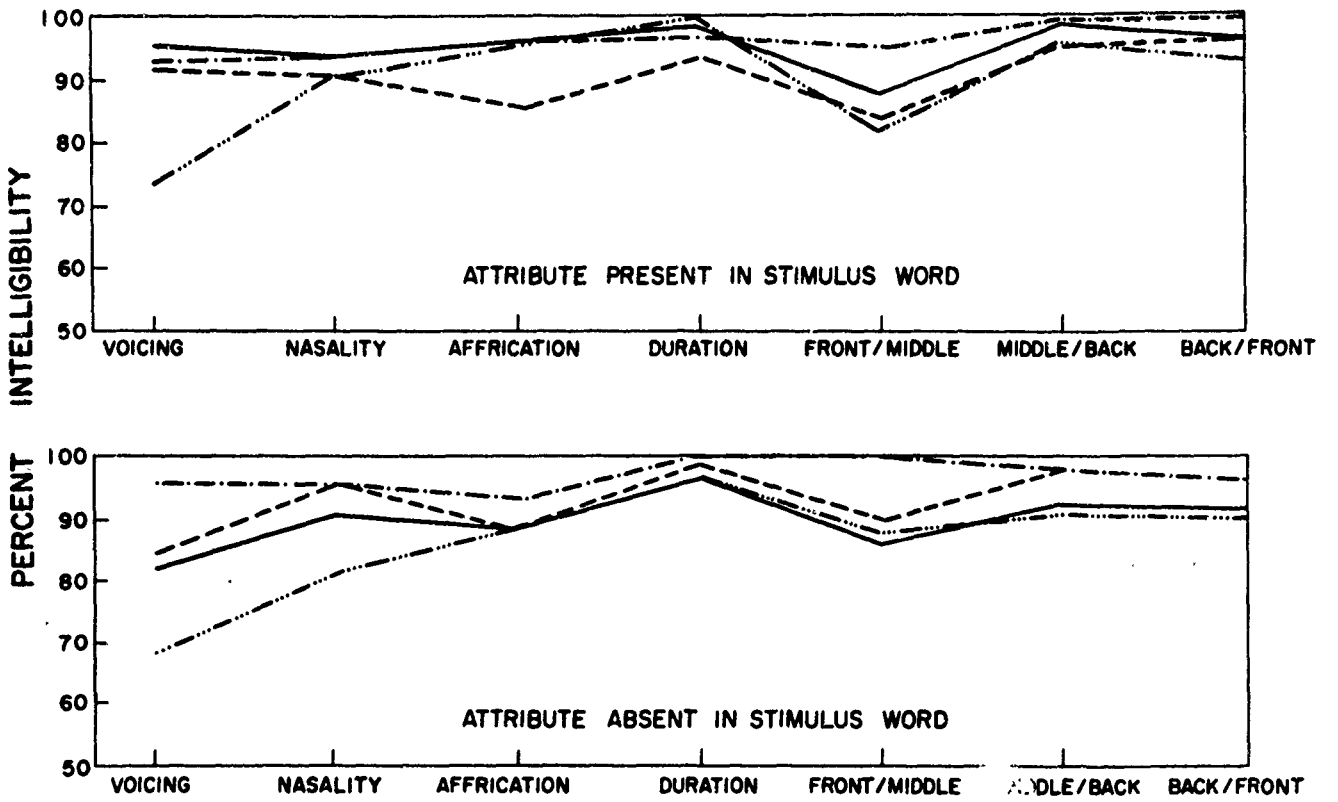


FIG. I-8a Diagnostic scores for three types of vocoderization.

Summary of Experimental Study No. I-9 Date: 5/17/65

Title: Evaluation of free-conversation tests of speech intelligibility.

Responsible Scientist(s): JM,MC,WV

Purpose:To determine the feasibility of diagnostically evaluating systems by means of presenting stimulus material conversationally rather than as words in isolation.

Methods & Materials:

Subjects: 8 University students (Group X)

Location: Tufts University

Stimulus Materials:A tape recording of a set of tasked directions in which nonsense syllables are incorporated as the stimulus materials. For this purpose, a nonsense syllable form of the DRT was developed.

Stimulus Conditions: 1) Band-passed from 200 Hz - 1500 Hz.
2) Band-passed from 1500 Hz - 4000 Hz.
3) Band-passed from 200 Hz - 4000 Hz.

Equipment:

Experimental Design: Subjects listened to the recording under each of the 3 conditions of frequency filtering. Response sheets, similar to those used for standard DRT evaluations,were used by the listeners.

Results & Discussion: From the results of this preliminary investigation, it appears that testing is sensitive to system differences and yields valid diagnostic scores. This experiment made evident several problems with this method, such as amount of time required.

Summary & Conclusions: Results indicate a need for further investigation of intelligibility testing using a method of conversation.

Summary of Experimental Study No. I-10 Date: 6/21/65

Title: A normative study of two Diagnostic Rhyme Tests.

Responsible Scientist(s): MC, WV

Purpose: To provide normative data for several speakers on the DRT and FRT.

Methods & Materials:

Subjects: 8 male university students (Group 3)

Location: SRRC

Stimulus Materials: DRT (random form, 4 speakers)
FRT (4 speakers)

Stimulus Conditions: Unprocessed Recordings

Equipment: Crown tape recorder
Scott Amplifier
PDR-8 matched earphones

Experimental Design:
Subjects listened to recordings of the DRT and FRT as spoken by 4 speakers.

Results & Discussion: See below.

Summary & Conclusions:

Results indicate that master tapes of the speakers involved yield "typical" scores for intelligibility on both the DRT and FRT.

Summary of Experimental Study Nos Q1, Q2 and Q3 Date: 4/17/65

Title: Derivation of a Standard Unit Variance Scale

Responsible Scientist(s): JM, WV

Purpose: To establish a standard scale, based on the Unit Variance method.

Methods & Materials:

Subjects: 8 males (Group 3)

Location: SRRC

Stimulus Materials: 15 sets of 10 sentences spoken by five speakers.

Stimulus Conditions: 15 pairs of 10 sentences spoken by five speakers were processed through six vocoders, four of which were standard vocoders. Stimuli presented binaurally.

Equipment: 2 channel Crown Recorder (SS-800), reference standard (built-in audio channel mixers, amplifiers), and 8 sets of PDR-8 permoflux earphones.

Experimental Design: The vocoder pairs were prepared using the following matrix:

	VOCODERS					
	B	E	L	A	F	C
vocoders B		1	9	11	14	6
E			4	13	7	10
L				2	12	15
A					5	8
F						3
C						

Pairs of vocoders from 1 to 15, in that order, were presented twice to the listeners. Vocoders L, A, F, and C are the standard vocoders. Vocoder scale values are based on 800 responses. The data were analyzed using the Unit Variance method. The unadjusted scale values for the four standard vocoders and three experiments were as follows:

	Exp.1	Exp.2	Exp.3	$\bar{X}$
A	.8746	.7679	.4940	.7122
C	.2344	-.3884	-.2015	-.1185
L	-.7183	-.1332	-.0293	-.2938
F	-1.3872	-.6383	-.4593	-.8284

ADJUSTED SCALE VALUES

	Exp.1	Exp.2	Exp.3
A	.5645	.7964	.7270
C	.1676	-.4435	-.3733
L	-.4230	-.1698	-.1009
F	-.8377	-.7114	-.7811

Summary & Conclusions: The mean for each vocoder, based on three independent experiments, represents the standard scale value for that vocoder. The coefficient of correlation between differences obtained for four standard vocoder scale values and differences obtained from direct comparison of vocoders is .9851. The obtained standard scale has the properties of transitivity and unidimensionality.

Summary of Experimental Study No. Q-4 Date: 11/24/64

Title: Vowel-to-Noise Ratio as a standard to evaluate preference of an unknown speech transmission system.
Responsible Scientist(s): J.M. Jr., WV

Purpose: To evaluate the possibility of Vowel-to-Noise Ratio as a standard in determining preference of vocoderized speech (experimental vocoders).

Methods & Materials:

Subjects: 8 male university students (Group 1)

Location: Tufts University

Stimulus Materials: 28 conversational sentences used in SRRC speaker recognition experiments were recorded by two speakers (neutral speaker and poor quality speaker). List of 28 sentences is in Appendix I.

Stimulus Conditions: 7 vowel-to-noise ratios were recorded on Ch. I of magnetic tape along with unprocessed speech materials. Speech materials on Ch. II were processed through one vocoder condition.

Equipment: 2 channel Magnecord tape recorder (Model 728), Scott amplifier (Type 296). and 8 sets of PDR-8 permoflux earphones.

Experimental Design:

	REPLICATIONS						
	1	2	3	4	5	6	7
U/N ratio 1							
2							
3							
4							
5							
6							
7							

Results & Discussion:
Means and Standard Deviations were obtained using Method of Least Squares. Psychometric functions were plotted for each speaker. The mean and standard derivation for the neutral speaker was +5.24 dB V/N ratio and $\pm$ 5.2 dB, and for the "poor quality" speaker was +4.70 dB V/N ratio $\pm$ 5 dB.

Summary & Conclusions:
These observations indicate that it is possible to compare directly any unknown transmission circuit to a well-defined variable standard. The use of untrained listeners to evaluate transmission quality of an unknown system was found to be advantageous.

Summary of Experimental Study No. Q-5 Date: 12/1/64

Title: Use of method of pair comparisons to evaluate speech quality of the polymodal vocoder modes.

Responsible Scientist(s): J.M. Jr., WV

Purpose: To evaluate preference of 5 analog modes of the polymodal vocoder under different listening conditions.

Methods & Materials:

Subjects: 8 male university students (Group 1)

Location: Tufts University

Stimulus Materials: 5 conversational sentences, used in SRRC speaker recognition experiments, were recorded by a formally trained speaker. List of 5 sentences is in List Q-1.

Stimulus Conditions: 5 modes of AFCRL polymodal vocoder were used to process speech materials, which were recorded on two channels of the magnetic tape to make pair comparison tests.

Equipment: 2 channel Magnecord tape recorder (Model 728), Scott amplifier (Type 296), and 8 sets of PDR-8 permoflux earphones.

Experimental Design: 10 vocoder pairs were presented twice: Order I and Order II at 72 dB SPL (7 dB SPL higher than the normal level-65 dB SPL). 5 vocoders used in every combination of direct comparisons yield the following matrix:

		VOCODERS				
		A	B	C	D	E
VOCODERS	A		AB	AC	AD	AE
	B			BC	BD	BE
	C				CD	CE
	D					DE
	E					

The lower part of the matrix was not used.

Results & Discussion: The data were analyzed using Thurstone's case V for pair comparisons. (See Guilford's Psychometric Methods for detailed outline of the method). Preference scale values were as follows: Vocoder A = .24; vocoder B = .25; vocoder E = .01; vocoder C = -.15; and vocoder D = -.36.

Summary & Conclusions: Five analog polymodal vocoder modes were evaluated for their relative preferences using method of pair comparisons. Presenting the stimulus materials at a higher listening level (7 dB SPL higher than the normal level used in other experiments) did not change the relative order in which these vocoders were preferred. Vocoder B, however, was preferred almost equally as well as vocoder A in this experiment. Previous experiment (No. 3) was at a normal listening level (65 dB SPL). In experiment No. 3 scale values were separated more between vocoders A and B.

List Q-1

SRRC Conversational Sentences Used In Vocoder Quality Judgement Experiments

1. Don't try to finish them before Tuesday.
2. He knows how to paddle a canoe.
3. There was oil spilled all over the road.
4. I think I'll eat in the cafeteria tomorrow.
5. The United Charity Fund exceeded its goal.

Summary of Experimental Study No. Q-6 Date: 12/3/65

Title: Use of method of pair comparisons to evaluate polymodal vocoder modes.

Responsible Scientist(s): J.M. Jr.,WV

Purpose: To evaluate preference of 5 analog modes of the polymodal vocoder under different listening conditions.

Methods & Materials:

Subjects: 8 male university students (Group 1)

Location: Tufts University

Stimulus Materials: Same as in Experimental Study No. Q-5

Stimulus Conditions: Same as in Experimental Study No. Q-5

Equipment: Same as in Experimental Study No. Q-5

Experimental Design: Same as in Experimental Study No. Q-5, except that the listening level was increased to 79 dB SPL (14 dB SPL higher than the normal level-65 dB SPL).

Results & Discussion: The data was analyzed using Thurstone's Case V for pair comparisons. (See Guilford's Psychometric Methods for detailed outline of the method). Preference scale values were as follows: Vocoder E= .30; vocoder B = .16; vocoder A = .01; vocoder D = -.08; and vocoder C = -.38.

Summary & Conclusions: Five analog polymodal vocoder modes were evaluated for their relative preferences using method of pair comparisons. The scale values among three experiments (No. 3, No. 7, and No. 8) indicate that changing the level of listening, the preference judgements also change. In other words, level of listening to vocoderized speech is important in establishing preference scales. In addition, vocoder outputs must be carefully monitored for the best effect.

Summary of Experimental Study No. Q-7 Date: 12/10/64

Title: Use of method of pair comparisons to evaluate polymodal vocoder modes presented through a loudspeaker.
Responsible Scientist(s): J.M. Jr,WV

Purpose: To evaluate preference of 5 analog modes of the polymodal vocoder using a different transducer-loudspeaker.

Methods & Materials:

Subjects: 8 male university students (Group 1)

Location: Tufts University

Stimulus Materials: Same as in Experimental Study No. Q-5.

Stimulus Conditions: Same as in Experimental Study No. Q-5.

Equipment: 2 channel Magnecord tape recorder (Model 728), Eico amplifier, (Type HF12A), and loudspeaker.

Experimental Design: Same as in Experimental Study No. Q-5, except that the listening level was increased 10 dB SPL above the normal level used for earphone listening (65 dB SPL).

sons.

Results & Discussion : The data were analyzed using Thurstone's Case V for pair comparisons. (See Guilford's Psychometric Methods for detailed outline of the method). Preference scale values were as follows: Vocoder B = .25; vocoder A = .14; vocoder E = .07; vocoder D = -.15; and vocoder C = -.32.

Summary & Conclusions: Five analog polymodal vocoder modes were evaluated for their relative preferences using a loudspeaker as a transducer. The scale values for vocoders indicate, as compared to previous experiments, that preference depends not only as the listening level changes, but also as the mode of transducer changes.

Summary of Experimental Study No. Q-8 Date: 12/10/

Title: Use of method of pair comparisons to evaluate intelligibility of polymodal vocoder modes.

Responsible Scientist(s): J.M. Jr.,WV

Purpose: To evaluate subjective intelligibility of 5 analog modes of the polymodal vocoder.

Methods & Materials:

Subjects: 8 male university students (Group 1)

Location: Tufts University

Stimulus Materials: Same as in Experimental Study No. Q-5.

Stimulus Conditions: Same as in Experimental Study No. Q-5.

Equipment: 2 channel Magnecord tape recorder (Model 728), Eico amplifier (Type HF12A), and 8 sets of PDR-8 permoflux earphones.

Experimental Design: Same as in Experimental Study No. Q-5, except that the stimulus materials were presented at 65 dB SPL, and the listeners were requested to indicate under which conditions the sentences were more intelligible.

Results & Discussion: The data were analyzed using Thurstone's Case V for pair comparisons (See Guilford's Psychometric Methods for detailed outline of the method). Relative intelligibility scale values were as follows: Vocoder A = .30; vocoder B = .12; vocoder C = -.03, vocoder E = -.08; and vocoder D = -.30.

Summary & Conclusions: Five analog polymodal vocoder modes were evaluated for their relative intelligibility using the method of pair comparisons. These scale values suggest a possibility that a vocoder of good judged quality does not necessarily have the best judged intelligibility.

Summary of Experimental Study No. Q-9 Date: 12/1/64

Title: Use of method of pair comparisons to evaluate speech naturalness of polymodal vocoder modes.

Responsible Scientist(s): J.M. Jr, W.V.

Purpose: To evaluate subjective naturalness of 5 analog modes of the polymodal vocoder.

Methods & Materials:

Subjects: 8 male university students (Group 1)

Location: Tufts University

Stimulus Materials: Same as in Experimental Study No. Q-5.

Stimulus Conditions: Same as in Experimental Study No. Q-5.

Equipment: 2 channel Magnecord tape recorder (Model 728), Eico amplifier (Type HF12A), and 8 sets of PDR-8 permoflux earphones.

Experimental Design: Same as in Experimental Study No. Q-5, except that the stimulus materials were presented at 65 dB SPL, and the listeners were requested to indicate which vocoders sounded more natural.

Results & Discussion: The data were analyzed using Thurstone's Case V for pair comparisons (See Guilford's Psychometric Methods for detailed outline of the method). Relative naturalness scale values were as follows: Vocoder A = .34; vocoder C = .03; vocoder E = -.06; vocoder B = -.13; and vocoder D = -.13.

Summary & Conclusions: Five analog polymodal vocoder modes were evaluated for their relative naturalness using the method of pair comparisons. These scale values suggest that subjective naturalness may be different from subjective intelligibility and quality.

Summary of Experimental Study No. Q-10 Date: 12/1/64

Title: Use of method of pair comparisons to evaluate speech quality of polymodal vocoder modes.
Responsible Scientist(s): J.M. Jr., W.V.

Purpose: To evaluate consistency of preference judgements of 5 analog modes of the polymodal vocoder.

Methods & Materials:

Subjects: 8 male university students (Group 1)

Location: Tufts University

Stimulus Materials: Same as in Experimental Study No. Q-5.

Stimulus Conditions: Same as in Experimental Study No. Q-5.

Equipment: 2 channel Magnecord tape recorder (Model 728), Eico amplifier (Type HF12A), and 8 sets of PDR-8 permoflux earphones.

Experimental Design: Same as in Experimental Study No. Q-5 using 65 dB SPL.

Results & Discussion: The data were analyzed using Thurstone's Case V for pair comparisons (See Guilford's Psychometric Methods for detailed outline of the method). Relative preference scale values were as follows: Vocoder A = .23; vocoder E = .01; vocoder C = .01; vocoder B = -.04; vocoder D = -.21.

Summary & Conclusions: Five analog polymodal vocoder modes were evaluated for their preference using the method of pair comparisons. The consistency of preference scales between Study No. 3 and Study No. 12 shows that listeners use the same criterion when they are presented with identical speech materials on two different occasions. The discrepancies found between vocoder B on both occasions may be due to the equipment changes.

Summary of Experimental Study No. Q-11 Date: 5/12/65

Title: Use of method of pair comparisons to evaluate speech quality of the polymodal vocoder modes.

Responsible Scientist(s): J.M. Jr., W.V.

Purpose: To evaluate preference of 5 analog modes and 5 digital modes of the polymodal vocoder.

Methods & Materials:

Subjects: 25 male university students (Groups 2A, 2B, 2C)

Location: Tufts University

Stimulus Materials: 10 sentences were recorded by five speakers (neutral, high-pitch, low-pitch, rough, and smooth voice characteristics. 10 random orders of 10 sentences and five speakers were prepared. List of 10 sets and 10 sentences in the set is in Appendix 2.

Stimulus Conditions:

Sentences were processed through 10 modes of the polymodal vocoder (5 analog modes and 5 digital modes). 45 vocoder pairs were presented to the listeners.

Equipment: 2 channel Crown tape recorder (Model SS-800), audio channel mixer, Scott amplifier (Type 296), and 10 pairs of PDR-8 permoflux earphones.

Experimental Design: Four complete and one incomplete matrices of vocoder pairs were presented to the listeners.

	ANALOG VOCODERS 1	2	3	4	5
ANALOG VOCODERS 1					
2					
3					
4					
5					

	DIGITAL VOCODERS 1	2	3	4	5
DIGITAL VOCODERS 1					
2					
3					
4					
5					

	DIGITAL VOCODERS 1	2	3	4	5
ANALOG VOCODERS 1					
2					
3					
4					
5					

	ANALOG VOCODERS 1	2	3	4	5
DIGITAL VOCODERS 1					
2					
3					
4					
5					

Results & Discussion:

The data were analyzed using UVS method. Vocoder scale values are in descending order.

Vocoder				
	B	2.6966	K	-.5896
	E	2.6732	A	-.9516
	D	2.2560	H	-1.7184
	C	1.6635	F	-2.5670
	L	-.3860	G	-3.0767

Summary & Conclusions:

Summary of Experimental Study No. Q-12 Date: 5/17/65

Title: Use of method of pair comparisons to evaluate speech quality of polymodal vocoder modes.
Responsible Scientist(s): J.M. Jr., W.V.

Purpose: To evaluate preference of 5 analog modes of the polymodal vocoder.

Methods & Materials:

Subjects: 8 male listeners (Group X)

Location: Hanscom Field

Stimulus Materials: 10 sets of 10 sentences and five speakers (See Study Q-10)

Stimulus Conditions: 5 analog modes of polymodal vocoder were made into 10 pair comparisons test.

Equipment: 2 channel Ampex tape recorder amplifier , and 8 sets of PDR-8 earphones.

Experimental Design: A matrix of 10 vocoders.

Results & Discussion: The data were analyzed using UVS method. Vocoder scale values are in descending order.

VOCODER	D	.7924
	E	.7591
	C	.6186
	B	-.0959
	A	-2.0742

Summary & Conclusions:

Summary of Experimental Study No. Q-13 Date: 6/23/65

Title: Use of method of pair comparisons to evaluate speech naturalness of the polymodal vocoder modes.
Responsible Scientist(s): J.M. Jr.,W.V.

Purpose: To evaluate speech naturalness of 5 analog modes and 5 digital modes of the polymodal vocoder.

Methods & Materials:

Subjects: 10 male listeners (Group 3)

Location: SRRC

Stimulus Materials: Same as in Study No. Q-10.

Stimulus Conditions: Same as in Study No. Q-10.

Equipment: 2 channel Crown tape recorder (Model SS-800), audio channel mixer, Scott amplifier (Type 296), and 10 sets of PDR-8 permoflux earphones.

Experimental Design: Same as in Study No. Q-10, except that instructions were changed to give naturalness preferences rather than quality preferences.

Results & Discussion: The data were analyzed using UVS method. Vocoder scale values are in descending order.

VOCODER	D	1.7032	H	-.3714
	C	1.1952	F	-.3754
	E	1.0713	A	-.5320
	B	.6640	L	-.8431
	K	-.1902	G	-2.0503

Summary & Conclusions:

Summary of Experimental Study No. Q-14 Date: 7/12/65

Title: Use of method of pair comparisons to evaluate subjective sentence intelligibilty of the polymodal vocoder modes.
Responsible Scientist(s): J.M. Jr.,W.V.

Purpose: To evaluate subjective sentence intelligibility of 5 analog modes and 5 digital modes of the polymodal vocoder.

Methods & Materials:

Subjects: 8 male listeners (Group 3)

Location: SRRC

Stimulus Materials: Same as in Study No. Q-10.

Stimulus Conditions: Same as in Study No. Q-10.

Equipment: Same as in Study No. Q-10, except that the listeners had to indicate sentence intelligibility, rather than quality.

Experimental Design:

Results & Discussion: The data were analyzed using UVS method. Vocoder scale values are in descending order.

VOCODER	D	1.5413	A	-.3472
	B	1.2272	H	-.4386
	C	1.2058	K	-.6748
	E	1.0810	F	-1.4500
	L	-.1298	G	-2.4179

Summary & Conclusions:

Summary of Experimental Study No. SR-1 Date: 6/15/64

Title: Information Structure of Voice Ratings

Responsible Scientist(s):WV

Purpose: To provide normative data on the implicit dimensionality and information per dimension of multi-dimensional voice rating.

Methods & Materials:

Subjects: Thirty-two Brandeis University students; 16 male and 16 female (Groups 1A, 1B, 1C and 1D).

Location: Brandeis University classroom

Stimulus Materials: Tape recordings of 24 every day sentences as recorded per 16 male speakers.

Stimulus Conditions:

Unprocessed speech presented by loudspeaker.

Equipment: Magnecord Tape recorder, Scott amplifier, Scott S-2 speaker.

Experimental Design: Four groups of eight listeners, four male and four female, rated each voice on two trials. Order of speaker presentation was reversed on second trial.

Results & Discussion: Four orthogonal dimensions of perceived variability among voices were revealed by the use of factor analysis. (See attached Tables). The pattern of factor wordings suggested labels for the four factors as follows: I Pitch-Magnitude, II-III Loudness-Animation, IV Clarity-Beauty, V Normality. Factor II-III was shown by subsequent research to represent a super imposition of two potentially independent dimensions of listener response to voices. Analysis of results for five item-pairs, each identified with a particular PAT indicated that a total of 4.72 bits of speaker identity information is contained in the means of multi-dimensional rating by a crew of eight listeners.

Table SR-1a Materials and Methods of the Normative Voice Rating Studies

	First Normative Study	Second Normative Study
Size of Speaker Sample	16 adult males	24 adult males (including 16 from Normative Study No. 1)
Stimulus Materials	24 "every day sentences"	16 "every day sentences"
Stimulus Presentation Rate	1 sentence/five seconds	1 sentence/four seconds
Manner of Stimulus Presentation	Scott S-2 loudspeaker	PDR-8 headphones (diotic)
Frequency-passband	60-7500 H_z	200-4000 H_z
Number of Rating Dimensions	24	16 (selected from original twenty-four from Normative Study No. 1)
Number of Rating Categories	7	9
Identification of Individual Speakers	Identifying phrase spoken by each speaker (e.g. "This is Speaker No. 7")	Identifying phrase spoken by announcer (e.g., "Now you will hear the voice of Speaker No. 7")
Rating Standard	"Subjective impression of the typical voice," formed on the basis of preliminary exposure to all voices	The voice of the announcer (a "neutral voice" selected on the basis of results of Normative Study No. 1)
Listeners	Sixteen male and sixteen female college students (Brandeis University.)	Thirty-two male college students (Tufts University)
Relevant Listener Experience	None	Average of approximately 12 hours as subjects for intelligibility tests

Summary of Experimental Study No. SR-2 Date: 4/21/65

Title: Further Investigation of the Information Structure of Voice Ratings

Responsible Scientist(s): WV

Purpose: To search for additional perceived acoustic traits.

Methods & Materials:

Subjects: 32 male university students (Groups 2A, 2B, 2C and 2D)

Location: Tufts University

Stimulus Materials: Recordings of 16 every day sentences by 24 G.A. speakers.

Stimulus Conditions: Unprocessed speech band passed at 200-4000 Hz and presented at approximately 85 dB. "Neutral" voice announced all speakers and served as standard.

Equipment:
Crown Recorder
Scott Amplifier
PDR-8 Headphones

Experimental Design: Stimulus materials presented twice (two speaker orders) to all listeners who performed in groups of eight. Response forms (item order) partially confounded with interaction of listeners and trials.

Results & Discussion (See attached tables.)
Five factors emerged: I. Pitch-Magnitude; II. Loudness-Roughness; III. Animation-Rate; IV. Clarity-Beauty and V. Normality. Greatest amounts of speaker identity information carried by first three factors. A total of 4.7 bits of speaker identity carried by five item-pairs selected for estimation of PAT values associated with five factorial dimensions.

Summary & Conclusions:
Total speaker identity information transmission not increased by combination of: relative rating procedure, 9 rating categories, simplified rating form and increased stimulus presentation rate. However, dimensionality of listener response is increased by this combination of experimental conditions.

Summary of Experimental Study No. SR-3 Date: 7/15/64

Title: Information Structure of Voice Ratings of Frequency Filtered Speech.

Responsible Scientist(s): WV, JM

Purpose: To identify the spectral correlates of perceived voice characteristics.

Methods & Materials:

Subjects: Seventy-two Tufts University students; 36 male 36 female (Group X)

Location: Tufts University

Stimulus Materials: Recordings of 24 everyday sentences by 16 G.A. speakers.

Stimulus Conditions: LP at 750, 1500 and 3000 Hz; HP at 750, 1500, and 3000 Hz;
BP at 60-750, 750-1500, 1500-3000 and 3000-6000 Hz.

Equipment: Magnecord Tape Recorder SKL variable electronic filter
Scott Amplifier
Scott S-2 Loudspeaker

Experimental Design: Four males and four females served under each experimental condition. All subjects made two sets of ratings following preliminary "practice" trial.

Results & Discussion: SKL probably did not provide sufficiently sharp cutoffs for present purposes, though some trends are apparent.

Summary & Conclusions: Greatest total amounts of speaker identity information contained in 750-1500 and 1500-3000 Hz ranges. There is some indication that high-passing increases information transmitted via the PAT, Animation-Rate. Greatest amount of Loudness-Roughness information transmitted via the middle range of the speech spectrum.

Summary of Experimental Study No. SR-4 Date: 3/18/65

Title: Comparative Evaluation of Three Experimental Vocoders from the Standpoint of Speaker Recognizability

Responsible Scientist(s): WV, JM and MC

Purpose: To evaluate the effects of three speech synthesis techniques (conventional, conv. with Spectrum Flattening, and conventional with Vocal Response Synthesizer) upon speaker identity information transmitted via five perceived acoustic traits.

Methods & Materials:

Subjects: Thirty-two males (Groups 2A, 2B, 2C and 2D)

Location: Tufts University

Stimulus Materials: Recordings of 24 everyday sentences by 16 G.A. male speakers.

Stimulus Conditions: Unprocessed speech materials and same materials as processed by three vocoders.

Equipment: Crown Tape Recorder
Scott Amplifier
Scott S-2 Speaker

Experimental Design:

2A	2B	2C	2D
Unproc. Speech	Conv. Voc. Speech	Conv. Voc. Speech with SF	Conv. Voc. Speech with VRS

Results & Discussion

Spectrum flattening results in an increase of the speaker identity information transmitted via five perceived acoustic traits.
(See following Table.)

Summary & Conclusions:

The Voice Rating Method is sensitive to differences among various modes of synthesizing vocoded speech.

Table SR-4a The Structure of Speaker Identity Information in Absolute Voice Ratings of Vocoded Speech

Condition	Speaker Identity Information ["C"$\bar{\bar{x}}\bar{\bar{x}}(8,n)$] in Bits							
	PAT I(2,20)[1] Pitch-Mac.	PAT II(8,23) Loud-Rough	PAT III(10,19) Anim.-Rate	PAT IV(12,15) Clar.-Beauty	PAT V(7,9) Normality	SUM	"C"$\bar{\bar{x}}\bar{\bar{x}}(8,5)$	R^5
A)Conv. Analog Vocoder with Spect. Flat.[2]	1.90	1.71	2.39	.40	1.17	7.57	5.33 (5)	2.24
B)Conv. Analog Vocoder with Voc. Res. Synth.[2]	1.43	2.08	2.02	.37	1.04	6.94	4.36 (3)	2.58
C)Conv. Analog Vocoder I[2]	1.55	1.30	2.04	.50	.60	5.99	3.27 (3)	2.72
D)Conv. Analog Vocoder II[3]	1.71	1.39	1.97	.00	.62	5.69	3.25 (2)	2.44
E)Conv. Analog Vocoder with Mono. Pitch[3]	.80	1.11	1.44	.00	.34	3.69	1.90 (2)	1.79
F)Conv. Analog[3] Vocoder in Whisper Mode	.90	1.47	1.51	.65	.64	5.17	2.62 (2)	2.55
Aver. for Cond. C and D	1.63	1.35	2.00	.25	.61	5.84	3.26	2.49
Unprocessed Speech[3]	2.07	1.78	2.12	.67	.71	7.35	4.62 (4)	2.73

1. Numbers in parentheses identify items from Speaker Rating Form A which were used to evaluate each PAT.
2. Stimulus materials presented by means of loudspeakers.
3. Stimulus material presented by means of headphones.
4. Dimensionality of listeners response to five PAT's
5. Redundancy = $\sum_{j=1}^{n}$ "C" $\bar{\bar{x}}x(8,j)$ - "C" $\bar{\bar{x}}\bar{\bar{x}}(8,5)$

APPENDIX IV

ABSTRACTS OF PAPERS PRESENTED AT THE SPRING, 1965 MEETING OF THE ACOUSTICAL SOCIETY OF AMERICA

"Communication System Evaluation From the Standpoint of Speaker Recognizability," William D. Voiers, Sperry Rand Research Center, Sudbury, Mass., and Brandeis University, Waltham, Mass. Procedures have been developed by means of which multi-dimensional voice rating data can be analyzed to evaluate the capacity of a communications system for perceptually useful information as to speaker identity. In addition to a gross measure of capacity for speaker identity information, these procedures yield measures of speaker identity information transmitted via selected perceptual dimensions or perceived acoustic traits as well as via various physical dimensions of the speech signal. Results for representative designs and operating modes are described. The research reported in this paper was sponsored in part by the Air Force Cambridge Research Laboratories, Office of Aerospace Research, under Contract AF 19 (628)-4195.

"Effects of Stimulus Presentation Rate Upon Intelligibility Test Scores", Marion F. Cohen, Sperry Rand Research Center, Sudbury, Mass. There is a need for standardization of intelligibility testing as it is used for evaluating communications systems. The purpose of this experiment was to evaluate the effects of different rates of stimulus presentation upon intelligibility scores. Stimulus materials provided by the Fairbanks Rhyme Test and the Diagnostic Rhyme Test were recorded several times with various time intervals between words. They were presented to listening crews under two different conditions: 1) bandpassed from 200 - 4000 cps, and 2) processed by an 18-channel vocoder. The data for each condition were analyzed to determine the effects of the various stimulus rates upon both the intelligibility scores and the standard error reliability of these scores. The research reported in this paper was sponsored in part by the Air Force Cambridge Research Laboratories, Office of Aerospace Research, under Contract AF 19 (628)-4195.

"Performance Evaluation of the Vocal Response Synthesizer," William D. Voiers, Sperry Rand Research Center, Sudbury, Mass., and C.P. Smith, Air Force Cambridge Research Laboratories, Bedford, Mass. The Vocal Response Synthesizer was evaluated from the standpoint of speech intelligibility, speech quality and speaker recognizability by tests which included the Diagnostic Rhyme Test, the Fairbanks Rhyme Test and a specially developed test of speaker recognizability. Synthesized speech was also evaluated after being successively processed from two to four times by the Vocal Response Synthesizer. Comparative data for a conventional vocoder and a voice-excited vocoder were also obtained. The research reported in this paper was sponsored by the Air Force Cambridge Research Laboratories, Office of Aerospace Research, under Contract AF 19 (628)-4195.

"A Diagnostic Rhyme Test for the Evaluation of Communications Systems," M.F. Cohen (nonmember), J. Mickunas (nonmember), J.F. Miller (nonmember), W.D. Voiers, Sperry Rand Research Center, Sudbury, Mass. A test for consonant articulation has been developed to provide a practical method of system evaluation with respect to seven "articulatory dimensions". It utilizes a pool of 128 rhyming word pairs, each designed to test for the transmission of a specific feature. Either word of each pair may serve as the stimulus. The listener's task is simply to identify which member of the pair has spoken. Any number of equivalent forms of the test may be generated by randomly varying the stimulus word. Successive administration of two or more equivalent forms is feasible as a means of obtaining any desired degree of score reliability. Four administrations can be accompliished in ten minutes. With a crew of eight listeners the standard errors of the various scores are of the order of one percentage point over the range from eighty to one hundred per cent articulation. In addition to a gross score for each feature, sub-scores for false alarms and detection failures are readily obtained.

"The Effects of Frequency Filtering Upon the Information Content and Structure of Voice Ratings," William D. Voiers and J.F. Miller, Sperry Rand Research Center, Sudbury, Mass. Using a multi-dimensional voice rating form, each of seven listening crews rated the voices of sixteen male speakers under a different frequency-pass condition. Results are presented in terms of total amount of speaker identity information transmitted under each condition. Also described are the effects of each condition upon the information transmitted via selected perceptual dimensions or perceived acoustic traits. The research reported in this paper was sponsored in part by the Air Force Cambridge Research Laboratories, Office of Aerospace Research, under Contract AF 19 (628)-4195.

"Preference Scaling of Vocoder Speech," J. Mickunas, Jr. (nonmember), Sperry Rand Research Center, Sudbury, Mass. A study was performed to evaluate preferences for vocoder processed speech. Tape recordings of conversational sentences were processed through five vocoders to provide stimulus materials. The method of pair comparisons was employed. Stimulus materials consisted of identical sentences. Each sentence was processed through a different vocoder. Distances between vocoders were calculated using a more realistic statistical model which is different from those previously employed for purposes of evaluating communication systems. The arcsine transformation of observed preference percentages was used in conjunction with analysis of variance to derive scale values which more nearly satisfy the requirements of a psychological distance function. Data are presented for five vocoder designs. The research reported in this paper was sponsored in part by the Air Force Cambridge Research Laboratories, Office of Aerospace Research under Contract AF 19 (628)-4195.

Security Classification

DOCUMENT CONTROL DATA - R&D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1 ORIGINATING ACTIVITY *(Corporate author)*
Sperry Rand Research Center
Sudbury, Mass

2a REPORT SECURITY CLASSIFICATION
Unclassified

2b GROUP
---

3 REPORT TITLE
Performance Evaluation of Speech Processing Devices

4 DESCRIPTIVE NOTES *(Type of report and inclusive dates)*
Final Report - Scientific (1 June 1964 - 31 July 1965)

5 AUTHOR(S) *(Last name, first name, initial)*
Voiers, William D.; Cohen, Marion; Mickunas, Juozas

6 REPORT DATE	7a TOTAL NO OF PAGES	7b NO OF REFS
31 July 1965	151	46

8a CONTRACT OR GRANT NO.
AF19(628)-4195

b PROJECT NO. Task No. 4610-02

c DOD Element No. 62405304

d DOD Subelement No. 674610

9a ORIGINATOR'S REPORT NUMBER(S)
SRRC-RR-65-94

9b OTHER REPORT NO(S) *(Any other numbers that may be assigned this report)*
AFCRL-65-826

10 AVAILABILITY/LIMITATION NOTICES Qualified requestors may obtain copies of this report from DDC. Other persons or organizations should apply to the Clearinghouse for Federal Scientific and Technical Information (CFSTI), Sills Building, 5285 Port Royal Road, Springfield, Virginia 22151.

11 SUPPLEMENTARY NOTES

12 SPONSORING MILITARY ACTIVITY
AF Cambridge Research Laboratories
Office of Aerospace Research, USAF
Bedford, Massachusetts

13 ABSTRACT This study is concerned with the development of improved methods of evaluating experimentally processed speech and, in turn, speech-processing devices and systems. Three bases of evaluation are dealt with in the study. These are: Intelligibility, Speaker Recognizability and Aesthetic Acceptability or Quality.

A two-choice diagnostic rhyme test for the transmission of consonant information has been developed. It yields a total intelligibility score plus diagnostic scores relating to the fidelity with which seven binary attributes of consonant phonemes are transmitted to the ear of the listener. These attributes are voicing, nasality, duration and affrication, i.e., front (as opposed to middle) middle (as opposed to back) and back (as opposed to front).

For treating the problem of speaker recognizability, procedures have been developed by means of which listeners' ratings of voices on various perceived acoustic traits can be analyzed to predict speaker recognizability under any given transmission condition.

The problem of evaluating the speech channel from the standpoint of quality is treated by means of the standard unit-variance method. In this method, speech as processed by four representative vocoder systems provides standards with which experimentally processed speech is compared by listeners. Listener response data are analyzed to yield a value representing the position of the experimental system on a standard unit-variance scale of aesthetic acceptability.

Results of evaluations of representative vocoders are presented for each of the three evaluation methods.

DD FORM 1 JAN 64 1473

Security Classification

14 KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Speech Evaluation System Evaluation Intelligibility Speaker Recognizability Speech Quality						

INSTRUCTIONS

1. **ORIGINATING ACTIVITY:** Enter the name and address of the contractor, subcontractor, grantee, Department of Defense activity or other organization *(corporate author)* issuing the report.

2a. **REPORT SECURITY CLASSIFICATION:** Enter the overall security classification of the report. Indicate whether "Restricted Data" is included. Marking is to be in accordance with appropriate security regulations.

2b. **GROUP:** Automatic downgrading is specified in DoD Directive 5200.10 and Armed Forces Industrial Manual. Enter the group number. Also, when applicable, show that optional markings have been used for Group 3 and Group 4 as authorized.

3. **REPORT TITLE:** Enter the complete report title in all capital letters. Titles in all cases should be unclassified. If a meaningful title cannot be selected without classification, show title classification in all capitals in parenthesis immediately following the title.

4. **DESCRIPTIVE NOTES:** If appropriate, enter the type of report, e.g., interim, progress, summary, annual, or final. Give the inclusive dates when a specific reporting period is covered.

5. **AUTHOR(S):** Enter the name(s) of author(s) as shown on or in the report. Enter last name, first name, middle initial. If military, show rank and branch of service. The name of the principal author is an absolute minimum requirement.

6. **REPORT DATE:** Enter the date of the report as day, month, year; or month, year. If more than one date appears on the report, use date of publication.

7a. **TOTAL NUMBER OF PAGES:** The total page count should follow normal pagination procedures, i.e., enter the number of pages containing information.

7b. **NUMBER OF REFERENCES:** Enter the total number of references cited in the report.

8a. **CONTRACT OR GRANT NUMBER:** If appropriate, enter the applicable number of the contract or grant under which the report was written.

8b, 8c, & 8d. **PROJECT NUMBER:** Enter the appropriate military department identification, such as project number, subproject number, system numbers, task number, etc.

9a. **ORIGINATOR'S REPORT NUMBER(S):** Enter the official report number by which the document will be identified and controlled by the originating activity. This number must be unique to this report.

9b. **OTHER REPORT NUMBER(S):** If the report has been assigned any other report numbers *(either by the originator or by the sponsor)*, also enter this number(s).

10. **AVAILABILITY/LIMITATION NOTICES:** Enter any limitations on further dissemination of the report, other than those imposed by security classification, using standard statements such as:

(1) "Qualified requesters may obtain copies of this report from DDC."

(2) "Foreign announcement and dissemination of this report by DDC is not authorized."

(3) "U. S. Government agencies may obtain copies of this report directly from DDC. Other qualified DDC users shall request through _____."

(4) "U. S. military agencies may obtain copies of this report directly from DDC. Other qualified users shall request through _____."

(5) "All distribution of this report is controlled. Qualified DDC users shall request through _____."

If the report has been furnished to the Office of Technical Services, Department of Commerce, for sale to the public, indicate this fact and enter the price, if known.

11. **SUPPLEMENTARY NOTES:** Use for additional explanatory notes.

12. **SPONSORING MILITARY ACTIVITY:** Enter the name of the departmental project office or laboratory sponsoring *(paying for)* the research and development. Include address.

13 **ABSTRACT:** Enter an abstract giving a brief and factual summary of the document indicative of the report, even though it may also appear elsewhere in the body of the technical report. If additional space is required, a continuation sheet shall be attached.

It is highly desirable that the abstract of classified reports be unclassified. Each paragraph of the abstract shall end with an indication of the military security classification of the information in the paragraph, represented as *(TS)*, *(S)*, *(C)*, or *(U)*.

There is no limitation on the length of the abstract. However, the suggested length is from 150 to 225 words.

14. **KEY WORDS:** Key words are technically meaningful terms or short phrases that characterize a report and may be used as index entries for cataloging the report. Key words must be selected so that no security classification is required. Identifiers, such as equipment model designation, trade name, military project code name, geographic location, may be used as key words but will be followed by an indication of technical context. The assignment of links, rules, and weights is optional.